Discerning Industrial Activity in N.E. England - Networks and Clusters through the Lens of the Internet

A Thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Business Administration

by
John R Williams
May, 2008

Acknowledgements

This thesis has been something of a journey which at times has occupied space in a wide range of areas and activities. Many of these areas have on occasions been at some considerable distance from my prior knowledge, experience or indeed comfort zone. However the journey has been made easier by many people who gave their time and support, a few of whom are gratefully acknowledged below.

My principal academic tutors - Prof. David Charles of the Business School at the University of Newcastle upon Tyne with particular reference to the wide landscape of industrial clusters and to the overall direction of the thesis and Prof. Dimitris Assimakopolous of Grenoble Ecole de Management for advice with networks and mapping.

For internet matters and programming support - particularly Patrick O'Keefe and Mark Radford of Maxsi Ltd. and Martyn Erya of Shadowfax Design for trying.

For company URL data freely given in the interests of research - Norma Foster at the North East Regional Portal, the North East Chamber of Commerce for similar records and to colleagues at Trends Business research Ltd for access to their own industrial databases for comparison purposes and to Nathan Pellow of Enabling Concepts Ltd who generously undertook the deployment of an anonymous survey on company website design.

Although running a company at the same time as doing a part time degree has meant there is no time to look at research during the working day, the company and the Board of NEMI Ltd and Fabriam have helped in other ways by giving me free access to a wide variety of tools, books, data sources, administrative support and in particular access to significant bandwidth.

Finally a working lifetime of experience provided by employers, clients, buyers, suppliers and a wide range of industry contacts who demonstrated sometimes unwittingly, the very nature of industrial clusters, networks and Communities of Practice in action.

But most of all to my wife for everything else.

John R Williams. Cullercoats, 2008

Abstract

Research on industrial activity has taken many forms over last century as both practitioners and academics alike have sought to gain an understanding of the drivers of the economy in any particular sphere of interest.

One facet of this research effort is that associated with the determination of industrial clusters or the particular groupings of economically linked firms and other networked organisations. The tools available to researchers in the field have improved with time but still rely heavily on sources of data gained from databases describing the firm’s activity and other fields of information. Many researchers in this field have noted the shortcomings of such data sources and in particular the indications of a firm’s activity gleaned from its SIC (Standard Industry Classification) code. SICs were originally introduced to enable authorities to understand the make up of basic manufacturing industry and then the supporting service sector but with the passage of time and the introduction of completely new forms of work it has been necessary to expand both the scope and resolution of SICs, a process that invariably lags behind the development of new industry.

In the last 15 years the rise of the internet and its vast capacity as an information source has opened up the possibility of gaining new insights into the activity of firms and possibly the ways in which they interact. This research is concerned with an investigation into the usefulness or otherwise, of the internet and the world wide web as serious adjuncts to conventional information sources for the study of industry in general and networks and clustering phenomena in particular.

The research looks first at the practicality of using a corpus of regional company URLs from which has been extracted all descriptive text from each firm’s website and which is then stored in a database. This database describing the firm’s activity, markets and connections can then be interrogated to gain an insight into industrial activity across the region. Such an insight is much richer in both detail and depth than that to be obtained from the rather coarse grained SIC approach and its accompanying brief text descriptors.

The second part of the research, again following the definitions associated with industrial clusters, is to look at the connections between firms and possible untraded interdependencies. Again this is done using publicly available data sources on the

internet by a combination of embedded links, visible connections to external references and web derived in-links, the latter being third parties who reference the firm on their own web site. It has thus been possible to both discern commonalities of firm activity and also to draw some of the visible connecting networks as a result of these investigations which leads to the interesting conclusion that, for the sectors studies in the North East of England at least, we are looking not so much at the clustering of artefacts *per se* but the clustering of ‘competencies’ in a wide range of sectors that share both common antecedents and current practice in forms of activity requiring strong engineering skills for the design and manufacture of complex structures operating in difficult or even hostile environments.

The conclusion overall is that the internet does indeed offer additional insights over more conventional forms of determining industrial clusters but that such insights, at the present time, should be regarded as an adjunct rather than a complete method of analysis in their own right. It is further noted that the tools available for internet search are continually evolving as is the take up and use of corporate websites by all sectors of industry and sizes of the firm. A logical conclusion of this process would be that in the not too distant future the prospects for acquiring information from web based sources will be significantly enhanced.

Contents

Acknowledgements 2

Abstract 3

Contents 5

List of Tables 9

List of Figures 9

Nomenclature and Glossary 12

Chapter 1. Introduction and Methodology 14

1.1 Conceptual Framework and Theoretical Approaches 14

1.2 Problem Statement 15

1.3 Why this Approach? 17

1.4 Expected Contribution to the Body of Knowledge 18

1.5 Expected Contribution to Practice. 19

1.6 Research Methodology 20

Chapter 2. Discerning Industrial Activity 26

2.1 Background 26

2.2 Supporting Theory 28

2.3 Towards A Definition of Industrial Clustering 31

2.4 Virtual Arrangements 41

2.5 Communities of Practice 44

2.6 Critical Observations 45

Chapter 3. Cluster Analysis Methods 50

3.1 Preamble 50

3.2 Analytical Approaches 51

3.3 Problems with data 58

3.4 Problems with SICs 59

3.5 Summary of cluster considerations. 64

Chapter 4. How can the web help in cluster research? – A few clues from the literature 66

4.1 Preamble 66

4.2 The development of Tools. 67

4.3 An Example in Context 69

4.4 Links 70

4.5 Extracting text from the Web ... 70
4.6 General Conclusions and the basis of a route forward ... 73
Chapter 5. Networks ... 75
5.1 Introduction ... 75
5.2 Types of Networks ... 77
5.3 Network thinking applied to the firm ... 87
5.4 Summary on Networks ... 90
Chapter 6. General Conclusions on Literature Survey ... 92
Chapter 7. Examination and Development of Tools and Techniques ... 97
7.1 Introduction and Basic Road Map ... 97
7.2 URL Scraper Program #8 ... 106
7.3 Findings ... 108
7.3.1 URL Scraper results ... 108
7.3.2. Conclusion and Further Work on URL Scraper ... 111
7.4 ‘Phantom’ ... 112
7.5 Full search of 4300 URLs using Phantom ... 116
7.5.1 Preliminary Remarks. ... 116
7.6 Output from customised Phantom program – a comparison ... 118
7.6.1 Basis ... 118
7.6.2 Method ... 120
7.6.3 Conclusions on test ... 122
7.7 Observed Benefits and Disadvantages ... 123
7.8 Conclusion on preliminary results ... 125
Chapter 8. Further Development of a Functioning URL Derived Information Database. ... 127
8.1 Introduction ... 127
8.2 Basis ... 127
8.3 URL Finder ... 128
8.3.1 Outline and Function ... 128
8.3.2 Results and problems ... 129
8.4 Additional company records ... 131
8.4.1 Additional Data ... 136
8.4.2 The Influence of Google ... 137
8.5 Conclusions ... 140

Chapter 9. Further Work using the Combined Database 142
9.1 Program and Data setup 142
9.2 Results 145
9.3 Interpretation 145
9.4 Comparative work 145
9.4.1 Keyword Count 145
9.4.2 Discussion on Test of keyword frequency for ‘hard to find’ activities. 146
9.5 Conclusions on Searches with Combined Database 149
Chapter 10. Research into Linkages by the Firm (I) 152
10.1 Introduction 152
10.2 Network thinking and the research question. 153
10.3 A look at web linkages. 154
10.3.1 Preamble. 154
10.3.2 Basic Requirements 157
10.3.3 Potential tools 158
10.3.4 Existing Tools 158
10.4 Existing Tools with Modifications 166
10.5 New Tools – the Need 167
10.5.1 Tracking of look-up history by Cookies 167
10.5.2 A Spidering program for Web links 170
10.5.3 Specification, design and program outline 172
10.6 Results of tests with Spider program 174
10.6.1 Preliminary investigations 174
10.6.2 Further investigation and initial results 179
10.7 Conclusions 184
Chapter 11. Research into Linkages by the Firm (II) 186
11.1 Examination of specialised groups of firms 186
11.1.1 Initial Results of Link Spidering 189
11.1.2 Preliminary Conclusions on initial Link finding 201
11.1.3 Other ways of finding Links. 203
11.1.4 Links by Inspection 210
11.1.5 The Acquisition of In-links for the Subsea Cohort 221
11.1.6 Combined linkage set for the Subsea Cohort 228
11.1.7 Results and Observations 232

11.1.8 Comparison with earlier work on a Subsea 'cluster' 236

11.2 Further investigations into the nature of the Network 248

11.2.1 Further examination of the Subsea Network using UCINET....................... 248

11.3 Geographic Location... 262

11.4 Conclusion of work on web derived linkages... 267

Chapter 12. Investigations into Overlaps Between Clusters.. 273

12.1 Preamble.. 273

12.2 Design of the Test 273

12.3 Commentary and Conclusions 277

Chapter 13. Summary, Conclusions from the Research and Recommendations for Further Work.. 280

References cited in the Text.. 297

Other works consulted but not cited.. 312

Appendices.. 349

Appendix 1. Citations containing the clause "Industrial Clusters"- by year. 350

Appendix 2. Project Plan Outline 351

Appendix 3. Clusters: A Variety of Definitions 352

Appendix 4 Website finder walkthrough... 353

Appendix 5 SIC Profiles 356

Appendix 6 Communities of Practice - Definitions... 358

Appendix 7. Network Analysis Spider - Program Description................................. 360

Appendix 8. Chronology of URL database build. 364

Appendix 9. Example of Tracking Chart for In-links.. 374

Appendix 10. Examples of Output from OCW Tracking Cookies 376

Appendix 11 Social Networking in Business 379

Appendix 12. Survey of Firms Web linking policies... 382

Appendix 13 Definition of SNA terms. 387

List of Tables

Table 1. SIC metrics by group 110
Table 2. % of sites indexed by 'Phantom' & URLScraper 114
Table 3. Comparison of TrendsSCAN and Phantom data 122
Table 4. Test of keyword frequency for 'hard to find' activities. 146
Table 5. Time to spider URL data by predominant sector 182
Table 6. Projected time to spider 1000 regional Links 183
Table 7 - Firm count by search word 198
Table 8 - No. of In-links for PR firms by search engine 204
Table 9 - Location of In-links 206
Table 10 - Centrality ranking for subsea links (all) 233
Table 11 - Centrality ranking for subsea (in-degree) 235
Table 12 - Comparison of authoritative nodes (Top 50) by Collaboration 240
Table 13 - Comparison of authoritative nodes (Top 50) All sources 243
Table 14 - Multiple Measures of Centrality 259
Table 15 - Degree of Involvement by Firms in other Sectors 276
Table 16 - Top 10 SIC by firm count 357

List of Figures

Figure 1. Basic Flow Diagram 21
Figure 2 - Porter's 'Diamond' 37
Figure 3. Process Flow Diagram 99
Figure 4 URL Scraper outline flowchart 107
Figure 5 - Output screen from Phantom Search Program 119
Figure 6 - Search Engine Output Example 130
Figure 7. Example of starting page for the RBKS program 144
Figure 8. URL links vs. Time (Linear projection) 182
Figure 9. URL links vs. Time (Power law projection) 183
Figure 10 - URL links vs. Time (Power law projection) 184
Figure 11 – Electronics (all) 190
Figure 12 – Electronics (detail) 190
Figure 13 – Environment (all) 191

Figure 14 – Environment (detail) 192
Figure 15 - Defence (all) 193
Figure 16 - Defence (detail) 193
Figure 17 - Offshore (all) 194
Figure 18 - Offshore (detail) 195
Figure 19 - Marine (all) 196
Figure 20 - Marine (detail) 196
Figure 21 - Knowledge Intensive Business Services (all) 199
Figure 22 - Knowledge Intensive Business Services (detail) 199
Figure 23 - 14 PR firms with in-links (detail) 208
Figure 24 – Subsea: Embedded hyperlinks and observed connections (Random Network) 215
Figure 25 – Subsea: Embedded hyperlinks and observed connections (Circular Network) 219
Figure 26 - Subsea: Embedded hyperlinks and observed connections (Uniform Network) 220
Figure 27 - Subsea: In-links (Random network) 225
Figure 28 - Subsea: In-links (Circular network) 226
Figure 29 - Subsea: In-links (Uniform network) 227
Figure 30 - Combined linkages (Random graph layout) 229
Figure 31 - Combined linkages (Circular graph layout) 230
Figure 32 - Combined linkages (Uniform graph layout) 231
Figure 33 - Subsea Collaborative Network (after Siedlock & Andriani, 2006) 238
Figure 34 - Subsea network (all by type) 251
Figure 35 - Type 1 Firms (Private Sector) 252
Figure 36 - Type 2,3,4 organisations 252
Figure 37 - Egonet; NOF 253
Figure 38 - Egonet; Argonautics 253
Figure 39 - Egonet; Rigzone 254
Figure 40 - Egonet; ata 254
Figure 41 - HiClus of Geo Distances; NOF 255
Figure 42 - HiClus of Geo Distances; RenewTeesValley 255
Figure 43 - Location of Subsea Activity, North East Region 265
Figure 44 - Subsea related organisations located in Tyne and Wear 266

Figure 45 - URL finder - basic flow diagram 355
Figure 46 - SIC Profile of OCW Database 356
Figure 47 - Network Analysis Spider – Program Flow Diagram............................ 363
Figure 48 - Survey - Introductory message.. 383
Figure 49 - User Response form 384
Figure 50 - Chart 1, user response 384
Figure 51 - Chart 2, User response... 385

Nomenclature and Glossary

(for items not elsewhere noted)

General.

RDA – Regional Development Agency

VAT – Value Added Tax, with reference to VAT registered businesses. For 2006-7 the qualifying threshold was £61000.

www.defra.gov.uk – UK Govt. department responsible for environmental protection.

Industry and the firm

A number of terms have been used throughout the text to describe various groupings of firms and other organisations. Generally, in descending order of size these are:

- Industry – A large group of firms contributing to a recognised industrial activity e.g. automotive
- Milieu – the surroundings or environment that influence organisations
- Sector – A defined and identifiable grouping of firms and other organisations generally engaged in a single overarching activity e.g. Oil and Gas
- Sub-sector – A part of a sector again identifiable as being part of a sector e.g. Sub-sea
- Segment – used here in the marketing sense as a group of firms having a common characteristic e.g. niche or specialised players in a sector
- Cohort – a group with some statistical similarities
- Group – Used with reference to a generally smaller number of firms not fitting any of the above terms.
- List – A flat listing of firms or other data, usually as input to a spreadsheet.

The above terms are subject to some overlap and authors referenced do not always follow the above definitions. The last four terms are used synonymously to some degree.

Web Related

URL – Uniform Resource Locator

BASE URL – The top level of a domain, e.g. www.nemi-cai.com.

CSV – Comma Separated Values. A format for storing data in a readable format. Values are separated by commas, Microsoft Excel will read CSV files. Used as .csv or *.csv in the text.

Dead Link – A link that returns no information or an error.

Flash - Multimedia technology developed by Macromedia to allow much interactivity to fit in a relatively small file size. Websites that use flash may be partly HTML or completely Flash.

HTML – Hypertext Mark up Language. The language that web pages are written in. Web pages are sent from the server as HTML files and interpreted by a web browser. HTML files are essentially text files.

Link – Abbreviation of Hyperlink. A graphical or textual item that serves as a link to another web page.

Store – a file or array for data storage.

URL – Universal Resource Location. The address of a website or webpage.

Spider – a tool that scans websites for specified text or code.

Metatag – Embedded or ‘hidden’ text string within a website, usually containing keyword descriptors of the subject site and which are accessed or ‘harvested’ by many search engines.

Chapter 1. Introduction and Methodology

This chapter outlines the philosophy, structure and approach to this thesis and which in turn sets out the context in which the research has taken place.

1.1 Conceptual Framework and Theoretical Approaches

The motivation for this research has arisen from a number of drivers. These include firstly the desire to more accurately understand the structure of industrial activity in a defined area and in particular that associated with the phenomena commonly referred to as clustering[1]. A good working definition of clustering in an industrial context is from Van den Berg, Braun and van Winden (2001, p. 187):

"The popular term cluster is most closely related to this local or regional dimension of networks. Most definitions share the notion of clusters as localised networks of specialised organisations, whose production processes are closely linked through the exchange of goods, services and/or knowledge",

Secondly it is acknowledged and as will be discussed later, that many of the tools and data available to the researcher have known shortcomings particularly those related to firm descriptive data and many commentators have given the opinion that such deficiencies can lead ultimately to significant errors and omissions particularly related to the discernment of such clustering activity.

In recent years the notion of industrial clustering has been subject to much research and investigation and the number of available publications on the subject is significant. Much of this research activity stems from the part that industrial clusters are thought to play in the competitiveness of defined areas, such concepts having being promoted on a world stage by prominent authors. This has also translated into encouragement by a variety of public authorities throughout the world charged with promoting economic development in their particular region of responsibility on the basis that an industrial cluster can be shown to be a 'good' device for increasing such

[1] The term **clusters** in an industrial context is discussed at length in the text. For illustration a number of working definitions by different authors are given in Appendix 3. Clusters: A Variety of Definitions.

economic activity and hence associated wealth creation within a defined geographical region.

There are many ways of discerning existing industrial clusters although many authorities charged with economic development are often interested in ‘potential’ or ‘emergent’ clusters on the basis that public intervention might aid or speed up the process of cluster development thus bringing some hoped for benefit to their region. Such potential clusters are more difficult to deal with in the sense that they are not as visible as more mature and clearly functioning clusters which have many readily observable industrial entities and supporting organisations.

A key to all research in this area therefore is an understanding of the industrial milieu in any particular locale. This requires accurate data on companies and related and supporting organisations and also an ability to discern the influence exerted by so called ‘untraded interdependencies’ an example of which might be a pool of skilled labour associated with the strength of the education and research elements of a local University.. It is also vital to have an understanding of the nature of the relationship between firms and such supporting entities but data describing such relationships are difficult to obtain and much depends on the skill and knowledge of the investigators. With regard to firm data from public and commercial databases, these have formed the basis of much cluster research for many years in spite of a number of known shortcomings. It is in relation to some of these deficiencies that the idea arose of augmenting the data in some way by the use of the biggest potential information resource of modern times – the Internet and the World Wide Web.

1.2 Problem Statement

In simple terms, it is difficult to find out what ‘goes on’ in companies, descriptive data is often not up to date, wholly accurate, relevant or rich enough in detail to differentiate between similar and same as far as activity is concerned. The relationships between companies and other supporting organisations have to be gleaned by direct means and in any case many companies do more than thing in more than one single market. We need to find a way of overcoming or at least mitigating these problem areas.

In the context of the above paragraph the research question is:

‘Can the use of the internet and the world wide web as an information resource add anything useful to more conventional methods of researching industrial clusters and networks?’

The general assumption here is that such a vast and powerful information source must be able to provide some additional insight into the basic description of individual firms that go to make up sectors, industries, networks and clusters.

However the internet is completely unregulated, full of ‘noise’ of various kinds as far as information is concerned and it may be that the effort required to extract and codify such potential new knowledge on a larger scale is not worth the struggle at the present time.

A perfectly legitimate outcome of the research therefore could be an answer of the form ‘shows promise but needs more work’ at least as far as the present state of development of the internet, associated search tools and the world wide web is concerned.

It is worth noting the issue of elapsed time for this work and the rapid pace of development of internet technologies and tools together with the associated diffusion of internet use on a world stage. This has been important in the sense that some tools not available at the commencement of the research were developed subsequently and constant revisiting of the basic schema was required to update what had been examined earlier.

The research question is generic but to keep the scale of the research manageable in terms of data sets it was carried out on a regional scale, the region being the North East England. Using such a bounded geographical area has some advantages in that the author has worked in the region for many years, has access to a number of appropriate data sources and has some understanding of the underlying structure and dynamics of regional industry. Additionally whilst such background knowledge is convenient, the North East is the smallest of the RDA[2] regions and this to a degree

[2] RDA – Regional Development Agency.

makes it more ‘manageable’ in data manipulation terms. However regarding the research question and proof of concept, the region and its population should be quite large enough to form a robust conclusion that can be scaled to a larger region or indeed a country. The disadvantage to having a defined geographical basis is that ‘connections’ outside the boundary are not recognised and in particular those near the border of the region may have connections with other firms only a short distance away but still outside the region. This is a problem that is recognised and discussed in the context of the research.

1.3 Why this Approach?

In the opening paragraphs, the notion of using the internet as an information resource in the context of supporting industrial cluster studies was briefly noted. As this is a fundamental precept of all the subsequent research the discussion is expanded in this section.

The rise of use of the internet coupled with the power of modern search engines, has given us a capability to quickly find information about every aspect of human endeavour, provided that such information has been published on the web. Such search capabilities include industrial entities and it does not need much of a leap of thought to realise that in the context of gleaning company information the internet is a potentially very powerful tool. Using a proprietary search engine such as Google it is possible to find the web site of firms or other organisations of interest on the basis of a few well chosen keywords. If this could be done on a large scale for a defined area then it should be possible to find out more about the structure of firm activity in a defined geographical area than might be determined from a SIC (Standard Industry Classification) based dataset alone.

The notion of combining say a database of a corpus of industrial firms within a defined boundary with relevant information gleaned from individual company websites is very seductive. It holds out the prospect of gaining descriptive knowledge on ‘hard to find’ company activities not captured by current SIC descriptors. An example might be those technology related trading activities that have only very recently come to prominence on the industrial scene and of which nanotechnology or bio-informatics would be examples. In addition to firm descriptive information there are other possibilities regarding the derivation of useful information to be obtained from internet based searching. In cluster research the whole notion of the vitality of a working cluster, to be discussed later, relies on communication of some sort between constituent members which implies some form of network. In web sites many companies reference links and also companies communicate electronically with their peers, buyers and suppliers and a whole host of supporting organisations. It may thus be possible to augment manual methods of cluster research in some way by finding digital evidence of trades in goods or knowledge.

The reality however of such an undertaking, on a scale large enough to be both statistically significant and geographically interesting is non trivial and rudimentary attempts to scope the effort required have indicated that there are significant problems, some of which may be intractable at the present time.

It should also be noted that the rate of development of technologies associated with the internet and their adoption on a global scale can be truly staggering when viewed from even a short term historical perspective. This leads to the not unreasonable assumption that should such rate of progress in either internet technologies or their take up continue then tools may become available that we cannot at present envisage. In some respects therefore this research has been a journey with a broad remit as discussed below and with the means of answering the research question being subject to modification and enhancement as the research progressed.

1.4 Expected Contribution to the Body of Knowledge

The use of the internet as an information resource to undertake or augment regional scale industrial cluster type studies is relatively new with few examples. Many potential users seeking to gain knowledge about companies by using the internet in a similar context are driven by the possibilities for commercial market opportunities rather than basic research and similarly they may also seek to gain knowledge on the firm as an aid to competitor analysis. A number of specialist companies, whose capabilities range from market research to Search Engine Optimisation (SEO) offer tools or consultancy in this field. Two examples amongst many are given in the footnote.[3] [4]

There are however a significant group of researchers [Chakrabarti *et al* (1999), Pierre (2001)] who are developing methods for large scale information retrieval. The researchers in this field are often from a computing science, artificial intelligence or machine learning background and are often following lines of enquiry primarily in one or more of these areas with the examination of some specific economic activity being undertaken only on an infrequent basis and usually to demonstrate the efficacy of some particular line of development rather than as an end in itself. One such line of research is described in Section 4.3 .

The research will answer the research question to some degree as that which is new is related to the discovery of the ways by which the internet, in all its many facets, can help to elicit knowledge about the firm in a cluster context together with interactions with other firms and the general industrial milieu within a defined region. The research will also point up the limitations associated with such an approach.

Given the evolving nature of the web the contribution to knowledge from the research at this time is a combination of what can be done now together with an informed prediction of what may be possible in the future.

1.5 Expected Contribution to Practice.

The emphasis here will be on the practical implementation of positive results.

It will be shown that the research effort has lead to the development of a methodology and possibly tools that can be used by practitioners to better discern and understand the pattern of industrial activity within a given geographic area. There is no doubt that practitioners in cluster research and related policy development face many unknowns regarding the accuracy and resolution of data and as a result such studies often contain as much art as science and any tools that can be developed that aid our understanding of the make up of industry in a defined region

[3] www.ultimathule.co.uk

[4] www.e-consulting.com

will be of help. In addition many modern industrial endeavours spawn technologies and hence firms whose activities outpace the ability of the various national tracking systems to keep pace with their appearance on the industrial scene. Thus a particularly desirable outcome of the research is to show that the methods developed can identify and cope with such 'new' activities on a timescale that is free of the present somewhat coarse grained approaches and associated limitations.

The contribution will thus be of the form that enhances our understanding of what can be done and also what cannot reasonably be undertaken thus highlighting where further development is required.

1.6 Research Methodology

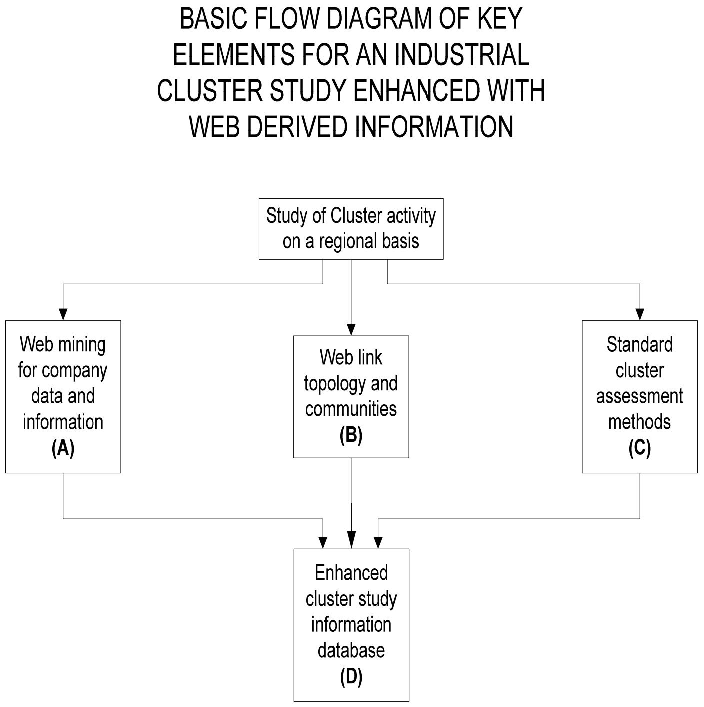
Outline

This thesis investigates the routes to new, internet based methods of discerning the profile of industrial activity in a given geographical area and compares the usefulness (if any) of this approach with more traditional methods of, *inter alia,* cluster analysis. There are three main routes to be investigated in the research within the aims as described above. It is intended that part of the fulfilment of these aims will be the building of an enhanced capability for the study of industrial clusters. This approach will have three elements as shown in Figure 1

These elements are all discussed at length in the body of the thesis but basically they are:

- 'Conventional' industrial cluster assessment methods (block C)
- The derivation of firm data, text and information from the websites of company's and other organisations (block A)
- An indication of the relationships that firms have one to another and to their environment through information gained from the web generally (block B)

Figure 1. Basic Flow Diagram


BASIC FLOW DIAGRAM OF KEY ELEMENTS FOR AN INDUSTRIAL CLUSTER STUDY ENHANCED WITH WEB DERIVED INFORMATION
Study of Cluster activity on a regional basis
Web mining for company data and information (A)
Web link topology and communities (B)
Standard cluster assessment methods (C)
Enhanced cluster study information database (D)

In this chapter, for each of these blocks there is a background setting the scene followed by a literature study appropriate to the research proposed. From preliminary work undertaken it appeared that there was little in the way of published work using the internet directly to augment an information database to be used primarily for the study of industrial clusters. On the other hand there is a mass of literature on related subjects where the authors have sought to research some other topic not directly related but which might be adapted or used here in a new way to support this research. The first section (block A) describes that part of the investigation whereby organisation descriptors derived from individual company web sites are used as the basis for discerning like minded activity and as a proxy for clustering. This part of the research therefore has required some understanding of:

- The notion of mapping of industrial activity in general.
- Methods of discerning clusters of activity

- The world wide web, search engines, crawlers, spiders and specialised software routines for accessing data and information from a large number of web sites.
- Database manipulation and lexical analysis.

In addition to the above main activities for finding what a company does further study has been carried out on how a company interacts with other organisations (which is a contributor to cluster activity in the round). This is block B and as it is being examined in the context of the internet such study has involved networks derived from the searching of links on any particular site on the basis that a site that links to other sites in some way regards them as part of its 'network' of activity or possibly, for small professional or knowledge based groups as part of a network of practice as in Duguid and Brown (2000, p.141).

The final section (block C) is what might be regarded as conventional cluster analysis using known methods and part of the purpose of this section is to provide comparative analyses when looking at the contribution of the first two sections. In subsequent research this part comprises both an introductory literature review on existing methods followed by experimental work using the results of research carried out by using the new methods of finding information about the firm using the internet.

There are a number of decision points regarding what should proceed sequentially and which activities could run in parallel. Clearly an emphasis on a literature survey is a fundamental precursor to the various sections and technologies that may be used. However there are other sections such as the assembly of a firm database that will be required and this was started at an early stage with continuous improvement as the research progressed.

Detail of the method of approach is shown in the form of a project plan with individual tasks and precedence's as shown in Appendix 2. Project Plan Outline. To expand the detail of the methodology followed the main activities are :

- Review of Industrial Clusters
- Web Mining and related technologies
- Research into Web Derived Linkages
- Integration, Interpretation and Conclusions

- Other Activity

These headings are expanded below.

Review of Industrial Clusters

This part forms the basis of Chapter 2 which examines the background to the definition and interest in the study of industrial clusters. The literature is rich and extensive with a high proportion of papers having been written in the last 15 years. To quantify this to some degree, a Google Scholar[5] search, in late 2006 elicited in total some 5510 references containing the clause "Industrial Clusters". With such a formidable corpus of papers available it was clearly necessary to be selective in order to gain an understanding of cluster development sufficient to bear upon the proposed research question. In addition to the theoretical constructs that support cluster studies there is also much useful work available regarding the more practical aspects of determining the present extent of different cluster forms and a good understanding of what is available is a particular requirement for research into how such methods might be improved or augmented.

It is known that more recently the whole concept of clusters and their usefulness has come in for critical review by a number of authors. The final part of Chapter 2 therefore attempts a 'state of the nation' summary of what is likely to be useful in the context of the focussed approach being proposed here and in particular will discern gaps in the literature or in practice within the context of the research question.

It is has been observed by a number of commentators of which Porter (1990) is but one authority that there are shortcomings in cluster analysis methods and also in the data used to support such studies. The examination of the phenomena of industrial clustering is therefore examined in a critical sense in Chapter 3 with a view to

[5] Using Google Scholar it is possible to search across many disciplines and sources: peer-reviewed papers, theses, books, abstracts and articles, from academic publishers, professional societies, preprint repositories, universities and other scholarly organisations. However it is important to acknowledge that there may be multiple citations of the same article as each Google Scholar search result represents a body of scholarly work which may include one or more related articles, or even multiple versions of one article. For example, a search result may consist of a group of articles including a preprint, a conference article, a journal article, and an anthology article, all of which are associated with a single research effort. These aspects are also discussed in Chapter 2. Discerning Industrial Activity

finding these gaps in the literature and the scope for improvement again in the context of the research question.

Web Mining and related technologies

This section covers a review of prior research undertaken in order to better understand how information might be extracted from the web. Most research with respect to web mining has taken place relatively recently and the volume of such literature tends to track the rise of internet use but again the review undertaken here is focussed on being adequate for an understanding of the basic principles and techniques. As with cluster literature this research is about using existing tools in a new way or the possible development of new tools and methods in relation to answering the research question. It is not an internet based project *per se.*

The literature regarding mostly existing work on data mining and extraction from information available on the web is covered in Chapter 4. A key part of the research and a requirement for the study of industrial firms in a collective sense is a database of firms and other organisations that may contribute to an industrial cluster. Such a database should be sufficiently large to be statistically robust when looking at the make up of industry. It will not however seek to cover all firms in the region as this would be unreasonable in terms of the time and effort required to build one as no such database presently exists or could easily be assembled particularly with respect to the acquisition of URLs for each firm. Preliminary indications would suggest that although data on the firm exists in many public and private/commercial databases, at the time of starting out (2003) the key field of company URL is present in only about 40%[6] of records for even the best of databases and such a field is blank for many others.

Research into Web Derived Linkages

A further major section on web derived linkages will be examined in this part with the literature review mostly in Chapter 5 with additional further research on inter-organisation linkages outlined in Chapter 10. Here the purpose has been to

[6] See Appendix 8. Chronology of URL database build. and also footnote 22 for remarks concerning derivation of this figure.

determine if it is possible to derive linkages between firms and other organisations using evidence of some sort of electronic ‘trace’. Although this kind of investigation on linkages is normally carried out manually by direct contact with organisations willing to take part it is very time consuming and any method of gaining information on firm associations in a networking context would be useful. This however is perhaps the most problematic part of the research and as such a variety of investigations have been pursued.

Lines of investigation have included the tracking of cookie information, html links on web pages, web page references to and from other sites, firms, client lists or people of interest in a network and hence in a cluster context.

For the web derived associations found, graphical representations either in network forms or as mathematical constructs such as centrality for popular links have also been examined. Such schema do of course make the significant assumption that web linkages are some sort of proxy for trading or other economic or knowledge based linkages. Such an assumption may not be entirely valid in all cases and the degree to which this assumption may be true is discussed.

Integration, Interpretation and Conclusions

This key section is where the separate investigations into the contributions that the internet can make to cluster research have been brought together into a useful whole. An important part of this section includes a comparison of methods used to support cluster studies carried out ‘conventionally’ compared with those examined using the additional tools and constructs found of use during the course of the investigation.

Many commentators, to be discussed later have given their opinion that cluster investigation involves a degree of art as well as science. The section on interpretation of findings and subsequent conclusion will therefore be important.

Other Activity

It was expected that during the research that there would be a requirement for ad hoc tools to be developed to aid the process of data or knowledge acquisition or manipulation although at the start of the research there was no clear indication what these might be. Some time was therefore allocated in the overall plan in Appendix 2. Project Plan Outline for such developments with the assumption that these would be an adjunct to the main lines of research and would proceed in parallel.

Chapter 2. Discerning Industrial Activity

2.1 Background

Industrial clustering – the tendency of firms who are linked in some way, usually by an economic relationship, to 'cluster' together in the same place - has a long pedigree. The topic of clustering has been the subject of debate throughout the 20th century but has attracted particular attention once again since the mid 1980s. This chapter will summarise the historical roots together with current thinking in the field but it is the intention to scope the depth of such a review to that which is necessary to inform subsequent research. The primary objective of this thesis is to answer the research question and to do this it is necessary to have an understanding of the how industrial clusters come into being, how they function in practice, how they are studied and the shortcomings of such methods of study. A comprehensive and all encompassing critical review of all contributors to the cluster debate would be a significant distraction given that following a discussion on industrial clusters it will then be necessary to further undertake a distillation of the main elements of both internet based text mining and network theory.

As was noted in the Introduction the literature is rich and extensive with a high proportion of papers having been written in the last 15 years. To quantify this to some degree, a Google Scholar search, in late 2006 elicited in total some 5510 references containing the clause "Industrial Clusters"and of these 1750 citations were associated with the year 2005 with 814 in the year 2006. The assumption is made that these could be the year of publication and examination of a small sample shows this to be a reasonable assumption. Although some of these search references might not be entirely appropriate (for example they could refer to clusters of computers in an industrial context) and the year can also refer to other year's citations contained within the abstract as indexed by Google, it is the author's view that the figures quoted are more likely an under representation of the worldwide effort undertaken on the subject because Google does not capture everything and in any case some papers (mostly pre internet era) may not be web accessible at all. A graph using the same source for the growth in such publications from 1970 to 2006 is shown in

Appendix 1. Citations containing the clause "Industrial Clusters"- by year. The background to the way Google Scholar indexes such references was noted in footnote 5 on page 23.

Industrial clusters refer to the tight connections that bind certain firms and industries together in various aspects of common behaviour such as geographic location, sources of innovation, shared suppliers and factors of production, and so forth. [Bergman and Feser (2000)]. Similarly, Rosenfeld [(1995), p.13] defines business clusters as a *"geographically bounded concentration of similar, related, or complementary businesses, with active channels for business transactions, communications and dialogue, that share specialised infrastructure, labour markets and services and are faced with common opportunities and threats".*

There are many other definitions and a selection of these, drawing upon Martin and Sunley (2001) are shown in Appendix 3. Clusters: A Variety of Definitions

Industry cluster concepts are not particularly recent and indeed date in various forms from at least the last century. Magee (2005) brought together a number of historical studies on industrial clusters and networks in England from 1750. However most commentators use as their starting point the work of Marshall (1890) who talked about 'industrial districts' and argued that once the process of local specialised industrial concentration had got underway it becomes cumulative and socialised in the locality in that *"The mysteries of trade become no mysteries; but are as it were in the air".* Marshall's 'discovery' of industrial districts arose primarily out of his empirical studies of the steel and textile industries, principally in the United Kingdom. Building on Adam Smith's (1776) recognition of the benefits of specialisation, Marshall established that within these industries, the greater the opportunities that existed to split up the production process, the greater the chances were that specialist firms would develop. In discussing the nature of production in Principles of Economics, Marshall saw the benefits of production accruing to the individual large firm as being differentiated between internal economies of scale and those arising to the industry as a whole, or external economies: *"We may divide the economies arising from an increase in the scale of production of any kind of goods, into two classes -firstly, those dependent on the general development of the industry; and, secondly, those dependent on the resources of the individual houses of business engaged in it, on their organisation and the efficiency of their management. We may* *call the former external economies and the latter internal economies".* [Marshall (1890), p.271]

The theories elaborated by Marshall are particularly relevant, not because he was the first to think about early clustering but because many more recent commentators have used Marshall as an underpinning construct for more sophisticated concepts and analyses.

As such, clusters have been 'rediscovered' in many regions of mature economies. Industrial districts can now be analysed as 'new' types of industrial productive systems, because they transcend the neo-classical notion of the firm. One therefore finds an increased awareness being placed on institutional capacity, governance and institutional change in regional development. Within the existing literature Becattini's (1990) definition (which probably is the most accepted within the Italian scientific community), of an industrial district is *'a socioterritorial entity which is characterised by the active presence of both a community of people and a population of firms in one naturally and historically bounded area. In the district – and unlike in other environments, such as the manufacturing town – the community and the firms tend, as it were, to merge'* (p.19).

2.2 Supporting Theory

The study of various types of industrial clusters benefits from being done so in the context of conventional theories of strategic management on the basis that such theories have much to say about how the firm is organised and its place in the market. This short section therefore briefly summaries the economic context in which value may be created as a basis for subsequent work on industrial clusters having the potential to open up new opportunities for wealth creation. An outline of such theories are, as appropriate in the context of the research question, taken up and elaborated on in the body of the text.

Whilst most of the early work on this topic had a geographic dimension we also have to take into account the issue of virtual clustering (see section 2.4) which, due to the enormous potential for the distortion of geographical constraints by modern communications technologies, usually but not exclusively through the medium of the internet, represent a significant challenge to conventional theories of how value is created [Amit and Zott, (2001)]

Value chain analysis

The work of Michael Porter is returned to later but in the context of this section Porter's value chain framework (1985) analyses value creation at firm level. Value chain analysis identifies the many discrete activities a firm performs in designing, producing, marketing, delivering and supporting its product or service. Value can be created by differentiation along every step of the value chain, through activities resulting in products and services that lower buyer's costs or raise the buyer's performance.

Schumpeterian Innovation

Schumpeter (1934) considers innovation as a source of value creation. According to Schumpeter, technology and novel combinations of resources are the foundations of new products and production methods. These in turn lead to the transformation of markets and industries and hence to economic development.

Resource Based view of the firm

The resource-based approach considers the firm as a group of resources and capabilities that may lead to value creation [Penrose, (1959); Wernerfelt, (1984); Prahalad and Hamel, (1990); Barney, (1991); Peteraf, (1993); Amit and Schoemaker, (1993)]. A firm's resources are valuable if and only if, they reduce a firm's costs or increase its revenues compared to the situation that would have existed if the firm did not possess those resources. A more recent approach, the dynamic capabilities approach [Teece, Pisano and Schuen, (1997)] explores how valuable resource positions are built and acquired over time.

Strategic Networks

The strategic network approach is relatively new but has important implications for cluster study and particularly in relation to the concept of virtual clustering as will be discussed in Section 2.4 when considering in particular knowledge based firms.

The strategic network approach focuses on the implications of stable inter-organisational ties which are strategically important to participating firms for the purpose of value creation. Strategic networks may take the form of strategic alliances, joint ventures, long term buyer supplier partnerships and other ties [Gulati, Nohria and Zaheer, (2000)]

There are several sources of value creation in strategic networks: trust, resources and capabilities (especially those of large suppliers and customers, economies of scale and scope, knowledge sharing, shortened time to market, enhanced transaction efficiency, reduced asymmetries of information and improved coordination between the firms involved in an alliance.

Similarly John Kay (1993) talks about the ‘architecture’ of the firm in relation to competitive advantage. This stems from the notion of ‘distinctive capabilities’ as a source of competitive advantage and he cited four generic headings under which such actions could occur. These areas are Innovation, Strategic Assets, Reputation and Architecture. It is the latter which is of interest here in that the term ‘architecture’ refers not to buildings but to the network of relational contracts with and around the firm. Firms may establish their relationships with and among their employees (internal architecture) and with their suppliers and customers (external architecture) or among a group of firms engaged in related or complementary activities. Such internal architecture also has some resonance with the concept of Communities of Practice as discussed in Section 2.5. The value of architecture rests in the capacity of organisations which establish it to create organisational knowledge and routines, to respond flexibly to changing circumstances and to establish easy and open exchanges of information. Each of these is capable of creating an asset for the firm – organisational knowledge which is more valuable than the sum of individual knowledge together with flexibility and responsiveness which extends to the institution as well as its members.

Although Kay was writing with reference to the firm there are parallels with various kinds of industrial clustering activity to be discussed later in that a geographically based cluster might be regarded as behaving as a large firm for the purposes of a functioning entity with a common and shared architecture.

Transaction Cost Economics

Transaction cost economics identifies transaction efficiency as a major source of value, as enhanced efficiency reduces costs. It suggests that value creation can derive from the attenuation of uncertainty, complexity and information asymmetry [Williamson, (1975)]. In more recent work [Amit and Zott, (2001)] have shown that the diffusion of the benefits of the digital economy impacts on firms and organisations by reducing coordination costs and transaction risks. In addition to decreasing these direct costs of economic transactions Information and Communication Technologies (ICT) also reduce indirect costs such as the costs of adverse selection and hold-up. However it should be noted that the preoccupation of transaction cost economics with efficiency may divert attention from other more fundamental sources of value such as innovation and the reconfiguration of resources. The theory also focuses on cost minimisation by single parties and the opportunities for joint value maximisation that this presents.

At this stage in the evolutionary process towards industrial clustering one of the fundamental connectors with all the above perspectives is that associated with proximity. The above concepts (with the possible exception of virtual arrangements) either directly or by implication depend to a greater or lesser degree on the presence of interacting entities in geographic space.

Porter's 1985 value chain analysis requires for its successful implementation a tight knit group of buyers, suppliers and supporting entities. With the latter day rise of ICT based collaborative working geographic constraints are not so crucial but when Porter was first researching these concepts geographic closeness was a reality amongst the firms under study. Similarly with Innovation theories, these have been grounded in observed studies that innovation diffuses over a bounded geographic area and certainly in 1934 when Schumpeter first propounded such theories place was crucially important.

For the resource based view of the firm again many of the resources a firm possesses to reduce its costs or increase revenues are based on the knowledge and capabilities of its workforce or some other fixed in place capital assets, the constraints of geography being evident.

Even transaction cost economics with its emphasis on the use of ICT is more concerned with efficiency amongst firms rather than removing distance effects as is discussed later.

Finally with strategic networks, as noted these are both internal and external and although as with Porter the position may alter with the adoption of advanced communication technologies, again at the time of research geographical location was a fact of industrial life leading to further work on the phenomena of clustering.

2.3 Towards A Definition of Industrial Clustering

The cluster concept has captured the imagination of active policymakers and the serious attention of scholars particularly in the last two decades. Because clustering behaviour is such a pervasive aspect of modern economies and global trade, it draws the attention of many different disciplines and benefits from their scholarship.

What follows is a summary of some of the main of lines of thought on the idea of industrial clusters, what they are, how they come into being, the different types and their place within the wider economic structure of a region.

As a cautionary note, as with many economic (and particularly management science) constructs, the study of the various cluster concepts is subject to the vagaries of fashion in the sense that, at a particular point in time, many practitioners see the whole notion of industrial clusters as being a good route to economic development in their particular geographical area of interest or responsibility. As such, many regional economic strategies have been defined on the basis of industrial cluster theory until some other 'magic bullet' appears and the attention of policy makers moves on. A clear example of this occurred in the North East of England where the 2001 Regional Economic Strategy (RES) produced by the Regional Development Agency (RDA) had many references to a variety of cluster based strategies and initiatives. By 2005 responsibilities had changed within the RDA and the new RES published covering the same broad issues for the region contained no main references to industrial clusters at all. This kind of variability in strategic view due to either the fashion of the times or the personal view of decision makers is a theme that is returned to later in the discussion.

The aim of this section is to establish enough of an understanding of key cluster concepts, the typology, their strengths and weaknesses in order to inform the subsequent research. It is not a comprehensive study of industrial clusters and given the significant literature base as noted in Appendix 1 it should be regarded as a summary.

External Economies

After Marshall a number of commentators looked at the effect of agglomeration and industrial location theory [Weber, (1929)] and Weber and Hoover, (1937)] although at this time clustering was still not a term much used. 'External economies', i.e. economic benefits not generated within the firm were propounded by Marshall as having a geographic basis, a concept that modern theorists have returned to many times.

Further work started to appear involving the distinction between urbanisation and localisation economies and clustering, in particular the proximity between firms, shared infrastructure, better availability of specialised facilities and reduced risk and uncertainty for aspiring entrepreneurs [Isard (1956), Lichtenburg (1960), Vernon (1960,) Carlino (1979)]. Rosenfeld (1995, p.20) cites 'tailored infrastructure' as a key advantage in regional development and propounds the use of a scale economy logic thus: *"As industry concentration increases, individual businesses benefit from the development of sophisticated institutional and physical infrastructure tailored to the needs of specific industry".* This is perhaps an important comment in that we need to be clear that there is a difference between generalised urbanisation economies and true industrial clustering.

The Innovation Environment

Industry clusters are also regarded as an important tool in policies related to national innovation systems (NIS) and as enterprises do not conduct business in isolation neither do they innovate in isolation [Benneworth and Charles (2001)]. The stimulus to innovate comes from a mix of competition, knowledge spillovers and market pressure. *"Innovation and the upgrading of productive capacity is seen as a dynamic social process that evolves most successfully in a network in which intensive interaction exists between those 'producing' and those 'purchasing and using' knowledge".* [Roelandt and den Hertog (1999) p.1].

Similarly Lundvall (1996) defines a NIS as *"the elements and relationships which interact in the production, diffusion and use of new and economically useful knowledge…. That are either located within or rooted inside the borders of a nation state".* It is also worth mentioning in the context of innovation the concept of the 'innovative milieu'. A milieu is defined by Maillat (1991, p113) as a *"complex which is capable of initiating a synergetic process…..an organization, a complex system made up of economic and technological interdependencies. . .a coherent whole in which a territorial production system, a technical culture and protagonists are linked".* Like research on industrial districts, literature on the milieu focuses on the specific nature and quality of transactions, alliances and partnerships between enterprises. The focus however is less on bilateral ties than the degree to which they support a collective environment for innovation [Malecki (1997)].

With respect to an innovative milieu, the main thing is to place the emphasis on the processes that allow a localised production system to generate new products, new technologies or new organisations - in short, to innovate. To do so, a distinction can be drawn between the hardware (material aspects of the production system) and the software (non-material and cognitive aspects). It is indeed thanks to the milieu's intellectual and cognitive capabilities that innovation is generated (c.f. the work of the GREMI group[7]), particularly as regards know-how and the ability to identify and formulate innovative projects [Crevoisier, (1993)]. These capabilities enable the milieu, on the one hand, to autonomously develop specific and differentiated resources and on the other, to perceive, identify and devise projects that enable the localised production system to be adapted and renewed.

Co-operative Competition

In co-operative competition theory the most competitive firms find ways to work together even though they may at times go head to head in the development of new products and the battle for markets. Modes of co-operation are based on trust, familial ties and tradition which is probably why they work best in cultures that can support such activities. Becattini (1989) talked about the ‘third Italy’ with a strong emphasis on the role of the cultural and historical background of the districts and he was the first to point out that a skill that appears abundant in a specific area may be scarce on the world market — for example, people who have been manufacturing clothes for centuries tend to possess a kind of “clothing culture and knowledge” that is of great significance. Thus, Becattini (1992) extended Marshall’s analysis of the purely economic effects of agglomeration to a broader perspective, to include the social, cultural and institutional foundations of local industrial growth. He also introduced the idea of “embeddedness” of the local industrial structure as a key analytical concept in understanding industrial districts. Further examples on the same theme are given in Asheim (1997), Park and Markusen (1995), Park (1997), [Isaksen (1997)] and [Heidenreich (1996)].

By way of further example, Doeringer and Terkla (1995) offer two circumstances in which cooperation among co-located firms can pay off. The first is when just-in-time (JIT) inventory and delivery systems are used. They cite the joint location choices of Japanese manufacturers and their suppliers, which is often necessary to make JIT truly work, as evidence of how cooperation drives regional industry clustering. The second example is a function of the speed and frequency of interactions between companies in a regional industry cluster. The more frequent

[7] http://www.unine.ch/irer/Gremi/accueil.htm

and rapid the interaction between suppliers, the more likely companies are to identify niche markets and new specialised products. They characterise such dynamics as 'collaboration economies' or *"the ability to participate in, and respond rapidly to, changing design and manufacturing practices among firms that buy and sell from one another"* [Doeringer and Terkla (1995) p. 182].

Path Dependence

Path dependence refers to the general notion that technological choices, even seemingly inefficient, inferior or suboptimal ones, can, for a variety of reasons, assume a dominant lead over alternatives which then become self reinforcing. Path dependence can have clear geographical implications by virtue of the fact that businesses, as a general rule cluster in space (however see discussion on virtual clusters in Section 2.4).

There are numerous examples to suggest that being the first mover can be critical to success [Arthur (1989), (1994); Krugman (1991), p.60] and all the elements for clustering, together with scale economies and externalities reinforce the early path generated lead. An example is given of the almost current universal use of the QWERTY keyboard.

The process of cumulative advance in regions whose industries have established a competitive lead in given markets has been described as an example of a 'lock-in effect' [Arthur (1989), (1990a), (1990b)]. In principle, the initial lead may be as much a result of luck or historical accident as by business acumen. Either way, particular 'locational clusters' may be able to establish a type of monopoly advantage over industries in other places. How likely or sustained such a process would be is an empirical matter [Krugman (1996)]. Bergman and Feser again cite Krugman (1991, p. 60) with respect to the carpet industry in Dalton, Georgia. From an economic geographers point of view, it was certainly by chance that tufting technology was essentially invented there. There was no carpet technology Institute at the local university, no cluster of carpet producers in the region, and no history of carpet making among local workers. Yet Dalton became a leader in carpet production (indeed, a locus for a carpet industry cluster), scale economies and externalities reinforced its lead and by 2004 80% of the world's carpet production was manufactured in Dalton[8]. Because technology can be path dependent, regional development trajectories can similarly become path dependent [see also Meyer-Stamer (1998)] and historical events such the one above suggest that being the first-mover can be critical to development success.

Rivalry and the Influence of Porter

No discussion on any type of business clustering activity would be complete without a mention of Michael Porter who is perhaps the most influential exponent of economic localisation. Much of Porter's thinking is rooted in the notion of rivalry amongst firms being one of the determinants of competitiveness and for such rivalry to function it requires firms in close geographic space to be involved. However Porter was by no means the first to consider this and the notion that rivalry amongst competing firms enhances geographic competitiveness has stemmed from studies by Becattini (1978, 1979) and White (1981) and which informed the basis of much of Porter's work (1980), (1985), (1990) although Becattini's seminal 1979 work raised the concept of the 'unit' of analysis as being neither a sector nor a particular industry but that of the industrial district. Basically, Porter adopts the neoclassical view that a competitive industry structure competing on the same playing field ensures continued pressure to upgrade technologies, minimise costs, to innovate and other yardsticks of competitive companies. Additionally, rivalry will likely be stronger amongst competing firms if they are geographically concentrated. In such a case the dimensions of competition increase as firms in the same region will compete not just for customers but for resources and other non-traded services such as labour, capital, political support and the like.

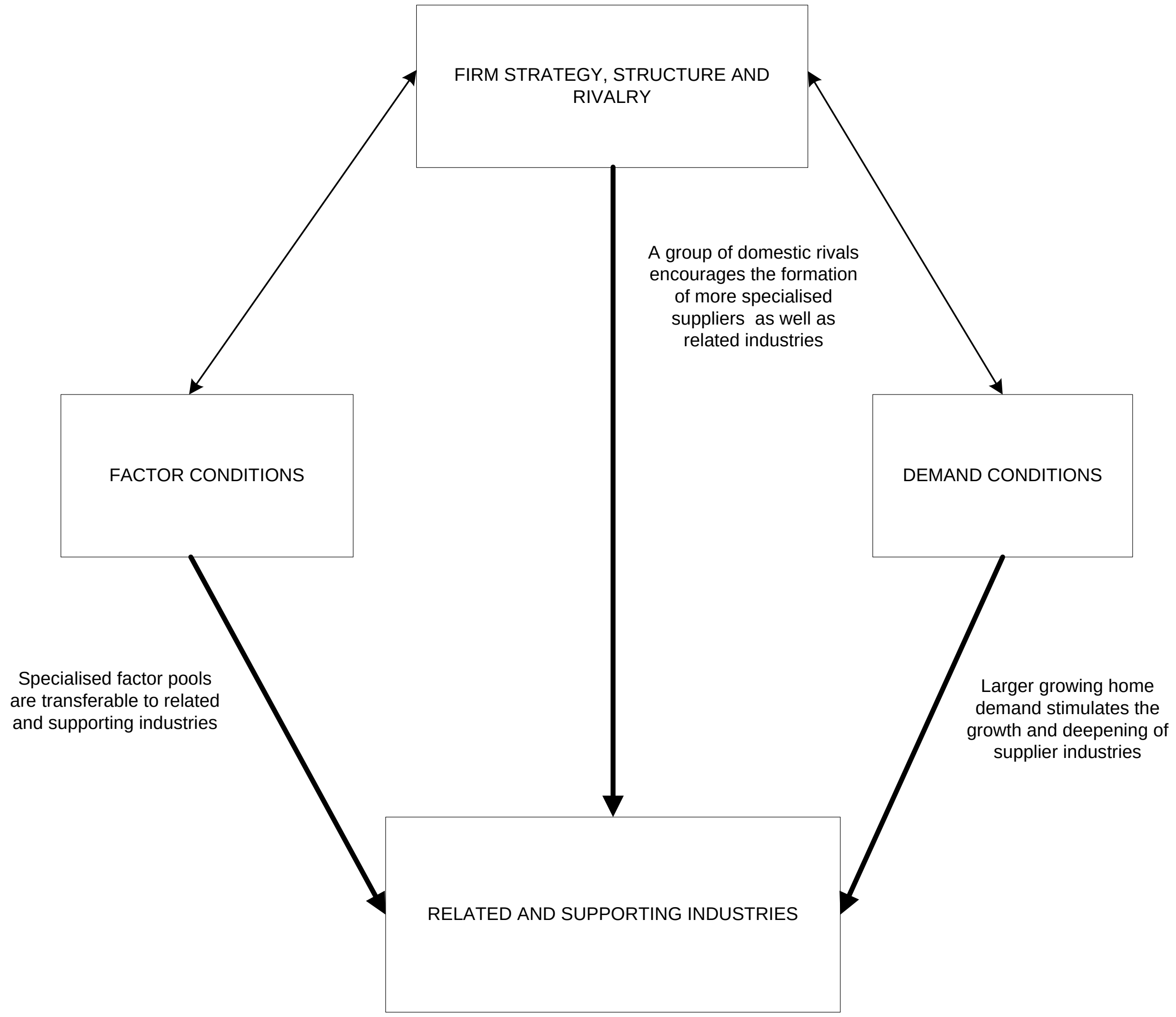
Porter's writings on industrial or business clusters thus rapidly become the standard concept in the field (certainly as far as many business schools were concerned) and Porter has promoted the idea of 'clusters' not just as an analytical concept but a key policy tool.

In his work on international competitiveness, Porter (1990) also argued that a nation's leading export firms are not isolated success stories but belong to a successful group of rivals within related industries. He termed these groups 'clusters', sets of firms and industries related by horizontal and vertical links of

[8] http://www.daltondirectcarpetinc.com/

various kinds including trading links. The significance of these industrial clusters resides in four sets of factors that constitute a ‘competitive diamond’. These factors are - firm strategy, structure and rivalry, factor input conditions and demand conditions together with related and supporting industries. The greater the intensity of interactions between the firms the greater the productivity of the firms concerned. These influences on the development of related and supporting industries are shown in diagrammatic form in Figure 2.

Figure 2 - Porter's 'Diamond'


FIRM STRATEGY, STRUCTURE AND RIVALRY
A group of domestic rivals encourages the formation of more specialised suppliers as well as related industries
FACTOR CONDITIONS
DEMAND CONDITIONS
Specialised factor pools are transferable to related and supporting industries
Larger growing home demand stimulates the growth and deepening of supplier industries
RELATED AND SUPPORTING INDUSTRIES

From this basic principle Porter then argued that the intensity of interaction within the ‘competitive diamond’ is enhanced if the firms in the cluster are geographically localised.

Porter has developed the concept considerably and used it as a policy tool related to a wide range of regional, national and international economies [Porter (1995, 1996, 1998, 1998b, 1998c, 2000, 2000b, 2001)] and particularly 1990 (pp149-159) regarding the clustering of competitive industries.

Thematic Clustering

Thematic clustering can take place when firms are concentrated around a theme such as for example - renewables. In this case a plethora of otherwise diverse firms related to environmental technologies, energy conservation, new forms of transport, clean energy, biomass etc might be organisationally bound to some degree even though their constituent parts might not be immediately thought of as having any common basis for co-operation.

As a less obvious example Mitra and Matlay (2000), in a paper on organisational learning in SMEs looked at the idea of 'connectivity', based on a review of the systemic view in the literature on innovation and industrial/business clusters, linking the two to demonstrate the relevance of organisational forms (clusters) to the innovation process. They then identified management considerations for organising appropriate competencies and leveraging firm-specific assets and outline management issues relating to externalities, innovation and the management of change - the essential elements of organisational learning. An 'Organisational Learning System' model was developed with a view to better understanding the innovation process of cluster-based SMEs. Further work took place identifying some of the key factors connecting firms in the high-technology circuit.

Other forms of Clustering

Clusters do not necessarily come into being along the lines of the categories noted so far. Often they grow through a process of ‘connectedness’. A 2005 study [Garnsey and Heffernan, (2005)] of clustering amongst hi-tech firms in the Cambridge (UK) area indicates that linkages have developed largely through the common origins of personal members of the cluster i.e. the University. This study was carried out by manually tracing the development of University spin-outs and the subsequent spin-outs from the original firms and so on.

Earlier but similar examples of this process are given in Chong Moon Lee *et al* (2000) in the ‘Silicon Valley Edge’ and in particular Chapters 8 and 10 of that book regarding the influence of the Fairchild Semiconductor company and Stanford University.

Clearly the tracing of such examples requires some effort on the part of the researcher and also willing access to and by the key players and other participants.

Apart from the effort involved, such a form of snowball sampling does rely for a comprehensive picture of the cluster, on a significant proportion of key figures taking part. If even a few of the principals, for whatever reason decline to participate, then it may not always be possible to plug these 'network gaps' by the knowledge of others in another part of the cluster.

An advantage of this approach should be that it may be possible to get a fine degree of granularity with respect to the number of firms and organisations that a key individual knows. The downside is that we only 'know who we know' and considerable art on the part of the researcher is required to gain a representative map of connectedness within the cluster, free from interviewee bias or lack of crucial knowledge.

The Issue of Untraded Interdependencies

Although not strictly a form of clustering the issue of what we mean by untraded interdependencies should be dealt with at an early stage as at a number of points in the thesis the term 'untraded interdependencies' will be used in the context of cluster definitions together with discussions on how such interdependencies might be identified. It will also be shown that the use of conventional based SIC methods gives difficulties to researchers when trying to identify identifying such contributors. Part of the problem would seem to be one of definition. The term 'untraded interdependencies' was introduced by Dosi (1988) and also used by Lundvall (1988, 1990) However Storper (1995, 1997) is perhaps best known for work in this area which refers to those cumulative-causation prone externalities which *"take the form of conventions, informal rules, and habits that coordinate economic actors under conditions of uncertainty. These relations constitute region-specific assets in production"* and "*of geographical differentiation in what is done, how it is done, and in the resulting wealth levels and growth rates of regions."* (Storper, 1995, p.5). In Storper's terminology therefore traded interdependencies are the formal transactions, the local input-output relations that take place between the economic agents in the region. Untraded interdependencies, on the other hand are the intangible assets of accumulated knowledge and localised learning of a geographical area that determines the direction of its development [Wibe (2003)]. An example might be firms that collectively contribute to the formation of a pool of trained labour experienced in the field. This argument seems fairly straightforward even if such untraded interdependencies are not easy to determine by any means. However as outlined above when Porter (1990) was talking in terms of the ‘Determinants of National Competitive Advantage’ in his book The Competitive Advantage of Nations one of these determinants of competitive advantage were ‘Factor Conditions’ of which one required element was the presence of ‘Related and Supporting Industries’. Porter’s view here is subtly different from Storper’s subsequent definition as it refers to firms or even sectors within a cluster that may be effectively non-trading but do in fact ‘support’ the competitiveness of the cluster. The picture is further blurred by a host of other ‘factor endowment’ conditions including knowledge resources and infrastructure both of which, when spread around the milieu might be regarded as an untraded interdependency. As having such a distinction in definition of ostensibly the same terminology is rather unsatisfactory, in subsequent use efforts been made to make clear the basis on which the term is being used. Generally the Storper definition refers to those untraded interdependencies found by inference whilst those found more by direct observation as ‘related and supporting industries’ fit more with the Porters ideas.

The Internet and Clusters.

Most clusters are deemed to run on information or knowledge exchange of one kind or another. The degree to which this happens may vary from a mainly manufacturing cluster with a smallish number of existing and well established networks to say a knowledge based cluster with many constantly evolving ‘digital trades’.

Clearly the capability of the internet to allow the rapid transfer of information without regard to boundaries should lead to new forms of ‘virtual clusters’. [Passiante and Elia (1999)] and [Formica and Mitra (2000)]. In these forms it would be expected that agglomeration should be easier to attain and that the ‘reach’ of the cluster for examples such as software production should not be at all limited by geographical considerations [Cairncross, F. (1997)]. To expand this example, there appear to be connections between software related industries in Singapore, Silicon Valley in California and Indian subcontinent hotspots such as Mumbai. This is partly because of the number of Indian origin entrepreneurs in Singapore and Silicon Valley with existing familial and trading ties and the ability to subcontract software development ‘over the wire’. However although there is anecdotal evidence, a number of news items and case studies of individual entrepreneurs operating this type of arrangement there are few published studies of the effect of this with respect to a fully developed cluster of any kind. This may be partly due to the necessity of having a 'control' cluster to measure the effect against a knowledge driven or virtual cluster. Such a task would be well nigh impossible because of the constantly evolving nature of this type of cluster.

Nevertheless by demonstrable actions if not by comparison, such models would seem to exist although the dimensions, scope and density of such clusters and networks are more difficult to measure than many more 'traditional' clusters as in Jurvetson (2000), Saxenian (1994, 2000), Castilla, Hwang, Granovetter E, Granovetter M, (2000).

In view of the rapid rise of use of the internet in all forms of knowledge exchange and also because of what is to come later in the research it is worth exploring further in the next section the topic of 'virtual clusters'.

2.4 Virtual Arrangements

In the majority of the cluster concepts considered thus far, geography or more precisely proximity has been shown to be a key factor. Porter, at least until 2000 has been a particularly strong proponent of this notion (but see Section 2.6 regarding some critical thoughts).

This 'organisational proximity' is the variable which is most sensitive to the dynamic interrelationship existing between economic organisations and the spatial change of an economy as suggested by Knox and Agnew (1998). Further, since everything happens in time and space, there may be a propensity to ascribe more causal importance to geographical or historical linkages than they really have for the simple reason that co-relations are more easily detectable in those dimensions.

The technological discontinuity generated by the Internet has therefore supported the emergence of a new form of organisational proximity that characterises a new competitive space, a notion that is not entirely inconsistent with Porter's view of clusters being driven by competitive forces. In this case however proximity is now not solely a spatial concept and 'how you do business' is more important than 'where you do business'.

Clearly with the emergence of the internet and the diffusion of cheap broadband connections, agents and groups may be proximate territorially or organisationally or institutionally or all of these together [De la Mothe and Paquet, (1998)].

In simple terms, with the advent of internet technologies, firms now have a far greater ‘reach’ than in the days when their horizons were limited to what and who they could see, hear, meet, and physically network with or become ‘involved’ through other local networks such as trade associations. With the availability of the internet the possibilities for networking are theoretically much greater than previously and with the added capability available through sophisticated search and retrieval tools again one would suppose that the emergence of clusters of complementary organisations would show a dramatic increase.

There are two difficulties. Firstly firms are run by people who ‘know what they know’ or more often by who they know and as trust is a key attribute of clusters [Woolthius (1999)] the ability to find potential cluster members on a bi or multilateral or project basis using the internet does not of itself make a cluster. The matter of trust between collaborating organisations in a cluster is usually only built up over a period of time. By contrast a network that comes together for a specific task such as a focussed project can be assembled relatively quickly. A comprehensive examination of this latter process particularly in respect of the role of personal networks in cross-national innovation in relation to the various EU Fourth Framework programmes has been looked at by Assimakopoulos (2007) in his monograph ‘Technological Communities and Networks, Ch. 4.

Secondly there is the problem of measurement. If proximity based clusters are difficult to find and measure in practice, such a task associated with virtual clusters, where members may be scattered across the world is truly daunting. The fact that it may be feasible to do so for very focussed networks, particularly those that can be traced via a coherent supply chain should not necessarily be used as a proxy for a ‘true’ cluster in the Porterian mould. A key element such as untraded interdependencies in a local cluster might involve some positive externalities due to agglomeration that are not as a result of trading relationships but which are driven more by local agglomeration effects an example being the establishment of a pool of knowledge and experience derived from the educational establishments in a locale, however an equivalent virtual cluster could use say a virtual learning environment (VLE) with educational or research content taken from anywhere in the world. Many examples given of this genre seem to be more networks than clusters, at least using one of the definitions in Appendix 3. Clusters: A Variety of Definitions

The notion of the ‘internetworked enterprise’ is discussed at length in Passiante, Elia and Massari (2003) in their book ‘Digital Innovation’. Here the authors argue that the opportunities provided by the new digital economy landscape are forging radical changes in the organisational structures of firms and even the creation of new forms of virtually clustered firms. Although these authors have coined the phrase ‘virtual clusters’ (VC) they are also referred to as e-business communities.

In such a virtual cluster each enterprise adds one or more distinct aspects of product/service value to the value of the network by exchanging digital knowledge with other members.

A corollary of the virtual cluster is the clustering of knowledge. This is not the same as clustering of knowledge based firms [Voyer (1997)] and numerous other examples in places such as Boston’s route 128 in the US, Catalonia (Spain), Lombardy (Italy) and Baden-Wuerttemberg (Germany) in which knowledge based firms are co-located. In this context it is as a VC defined above with the knowledge available as it were, at any place on the planet that can use the internet to link for a common purpose. Software industries seem particularly suited to this type of arrangement.

The study of virtual clusters is relatively new and many of the problems that arise in studying clusters in general i.e. ‘conventional’ proximity based clusters would appear to be exacerbated in the virtual world. For example addressing the question of the difference between networks and clusters and indeed when does a very rich networked environment graduate to being a cluster. This question is much more difficult to answer when dealing with internetworked enterprises because of the almost limitless connections that can be made at least at some level. This point comes up later in the thesis.

To try and summarise the different view points on virtual clusters some authors have taken the view that certain defined industrial activities that can be performed in an agglomerated economy with all the hallmarks of a cluster according to one of the definitions previously discussed, can be undertaken perfectly well irrespective of distance using modern telecommunication technologies. Critics of this view point out that the subtleties of proximity based clustering involving high levels of built up trust, norms and values shared and real and visible weak links and untraded interdependencies are just not present to the same extent in a ‘virtual cluster’. Even Alfred Marshall had pointed out how in clusters *“social forces co-operate with* *economic ones".* They further argue that it is also entirely possible for firms and others to be part of a proximity based cluster without actually having direct links with any of the members. These firms benefit from the general 'buzz' and the spin-offs available from co-location such as improved services, inward migration of skilled labour and the associated dynamism imbued in the local economy as a result of the cluster.

Virtual clustering phenomena such as have been studied seem for the most part to be highly selective, usually involving some aspect of ICT which lends itself to internet based transactional activity but it is the author's view that at the present state of implementation of these examples they are more sophisticated networks than true clusters.

2.5 Communities of Practice

Communities of Practice (CoP) were mentioned briefly in the Introductory work and it is appropriate to include the idea of a CoP under the section on Discerning Industrial Activity. As with the cluster concept a number of commentators have a view on what constitutes a CoP and for completeness most of these are shown in Appendix 6 Communities of Practice - Definitions although perhaps the two best known monographs on the subject are those by Wegner (1998) and Duguid and Brown (2000)

A common thread that seems to run through most of these is that they involve groups of people who work together to solve common problems. Johnson-Lenz, P & T (2000) for example define it as *"a group of professionals, informally bound to one another through exposure to a common class of problems, common pursuit of solutions, and thereby themselves embodying a store of knowledge."*

Another perhaps appropriate and more practical definition in the context of the research question is from Snyder and Briggs (2003 p.7) whose view is that **"***Communities of practice steward the knowledge assets of organisations and society".*

The inference from all the commentators seems to be that CoPs have a mostly small membership compared to say a sector or significant segment of industry and they can exist not only amongst separate groups but also within a single organisation.

In the context of our understanding of how communities work together it is probably fair to speculate that the very earliest artisan type of clusters such as that of the 15th Century printing press industry in Heidelburg, would be largely indistinguishable from one of the above definitions of CoPs. For large CoPs, small clusters or networks, at a certain point there is overlap not just in definition but in practice also. Duguid and Brown (2000 Ch. 6) also talk about networks of practice as equivalent to clusters. At this point in the literature reviews it was therefore felt appropriate that a CoP should be treated as a special case of a clustering.

2.6 Critical Observations.

In recent years there has been a serious questioning of some of the basic precepts of industrial clusters framed in the simple question 'where is the proof that they do any good?' Up to the point of such dissenting voices gaining some momentum many practitioners at least have taken as read cluster concepts because they seem to 'fit' with what may be happening on the ground, they seem to make sense and also because some high profile academics have actively promoted such concepts through articles in the management press, the conference circuit and national consultancy assignments. This section attempts to summarise some of the areas of doubt and challenge to the robustness of the subject and which comes from a body of opinion often from a background in economic geography. This is not a case of refining work already done so much as a serious questioning of the fundamental assumptions and the need for hard evidence that clusters do really exist and more particularly do they do any good in an economic development sense? This process could of course be seen as knowledge development and as more work is done by both academics and practitioners knowledge increases and the debate moves on. In particular, notions that at the time seemed rational and sound are revisited in the light of new evidence and some accepted theories supported by not much more than scholarly reputation or a fashion of the times are questioned regarding their validity. Such of course is the whole nature of research in general but in the case of industrial cluster research and application there seems particular opportunity for theory and implementation to run ahead of sound evidence that clusters both exist and make a positive difference to industry and economy. Although the debate amongst those studying the idea of industrial clustering and the various related issues has grown in intensity over the last decade or so there would still appear to be a dearth of research results clearly demonstrating that clusters do actually make a difference to a defined economy.

For example Sternberg (1996) points out that, so far at least, very few empirical examples of locally confined clusters have been quantified. This is partly because of the difficulty of measuring the pre and post event economic characteristics of a specific geographic location where the 'event' is a cluster coming into play. Even for a possible longitudinal study using say input-output tables, the presence of so many interrelated factors, the underlying untraded interdependencies and a whole spectrum of economic 'white noise' would tend to make such a task formidable. It is partly for this reasons that commentators have tended to look at clusters as at a point in time in order to demonstrate some supposed continuing worth.

In a clearer example Simmie and Sennet (1999) in a study of existing clusters and innovation in the South East of England found that in the case of some of the most innovative core metropolitan regions, co-location was a more frequently found condition than highly developed regular networking. They further went on to say that their data tended to support the proposition that innovative firms in core metropolitan areas like London, are gathered together there not so much because they need or use strong intra-industry networks or linkages, but rather because they are making use of the multiple pick-and-mix possibilities provided by the urbanisation effects of these large urban conglomerations. This goes back to considerations of agglomeration to minimise transaction costs versus more general externalities.

In addition to the above researchers have started to question some of the basic definitions and precepts being used. Perry (1999) for example contends that the definitional incompleteness of the cluster concept has been an important reason for its popularity. Steiner (1998, p.1) also noted that clusters have *"the discreet charm of objects of desire"*. In addition policy makers, particularly at the regional level have seized upon the cluster concept to promote public intervention in areas of industry that are felt to be desirable. We thus hear of terms such as 'aspirational,' 'emergent' and 'potential' as prefixes to the term 'clusters'. If one combines these with 'traditional' or 'mature' clusters virtually the whole of industry and commerce in any particular locale can be thus captured.

In particular Martin & Sunley (2001) have carried out a comprehensive critique of the various concepts and applications (or lack of these) and also have singled out

Porter for particular attention in this regard. Their contention, apart from the general one above regarding lack of hard evidence, is that there are three basic areas of query where the robustness of the arguments should be questioned.

Firstly Porter has rooted and promoted his cluster concept with a focus on the determinants of 'competitiveness'. This strikes a chord with policy makers keen to see their province of responsibility make progress on a world stage. The work of economic geographers in the same area has tended to be less focussed and they have looked more at agglomeration and industrial localisation.

A second but related issue is one of style and presentation and Porter's undoubted celebrated international status as a leading writer on business strategy. Much of Porter's work appeals to practitioners because of the self confident and authoritative style which lends his cluster concept an apparent authenticity and legitimacy that policy makers have found difficult to resist. In contrast economic geographers for example [Markusen, (1998); Martin, (1999); Glasmeir, (2000)], have much less influence on business policy and have been, by comparison, barely visible in the research agenda for active policy makers and industry champions. A possible reason for this is that their approach is perhaps more analytical than normative with a healthy scepticism for concepts that are so difficult to measure.

Thirdly, the very nature of the cluster concept itself carries confusion within it. Porter's cluster metaphor is highly generic in nature and semmingly deliberately vague and sufficiently indeterminate as to admit a very wide spectrum of groupings and specialisations. Although such definitional elasticity can be seen as a positive feature in that it permits a range of cases and interpretations to be included, Martin and Sunley regard this as problematic. They conclude that *"the concept has acquired such a variety of uses that it has, in many respects, become a 'chaotic concept', in the sense of conflating and equating quite different types, processes and spatial scales of economic localisation under a single, all-embracing universalistic notion"*.

In addition to definitional problems the nature of the methods used to determine the cluster in any particular locale have been subject to some scrutiny and for example, Enright (1996) described Porter's analytical methods as *"opaque at best"*.

One of the results of the above is that because of the seductive appeal of Porter's work, policy makers usually encounter a business or industry flavoured approach to identifying and analysing clusters that Bergman and Feser (1999) refer to as micro.

In such micro-level applications clusters are defined as a group of firms that produce similar products (i.e., industries) but which also hold key complementary informal and formal ties. These clusters may include some limited supply chain firm characteristics but in such studies explaining value-chains is less important than characterising ties between similar producers.

In many cases single industry clusters set the standard for studies under consideration by development policy officials who face a large portfolio of often very different interacting industries. This approach restricts its view to a single visible collection of similar sector firms, thereby overlooking linkages that some of its members may have with regionally co-located firms from very different sectors or the robust clustering of other sectors. A micro-level study then tends to document one cluster per region, usually that of its policy client. As a consequence there is apparent indifference to the presence of additional clusters, particularly those based on alternate criteria or detectable only from a wider spatial view or from data-intensive sources. An implication is that significant instances of region-wide industrial clustering go unrecognised by micro studies. At the same time, the labour intensive method of study all but precludes a region-wide investigation of all industrial clusters that might form the basis for "*seeing regional economies whole.*" [Bergman and Feser (1999) Ch. 3, p.3] This latter issue of avoiding an holistic view of possible multiple clusters within a region is a point that is of interest in the context of the research question in that, as will be shown, internet derived information is no respecter of geographic boundaries, sectors, expert input or anything else that requires a significant effort to track connections between firms.

As a final note on this section, the work by Malmberg and Power (2006) suggest that the whole cluster debate has become too idealised at the expense of practical utility. In their words *"we have suggested that the state of play in cluster research is currently rather unsatisfying. We have a situation where considerable conceptual confusion reigns and this confusion presents academics and practitioners with real 'cluster headaches'. In particular, we have pointed to the idea that the cluster literature has become far too concerned with conceptual Puritanism that aims to identify certain attributes and qualities that earmark a set of industrial activities as a 'cluster'. Indeed we have suggested that an implicit set of ideal types and an implicit model of cluster competitiveness have emerged to dominate many aspects of the cluster debate".* These authors also go on to assert that the available evidence suggests that, if we are interested in knowledge creation and knowledge-based innovation, localised clusters seldom appear to be the localised systems of interrelated firms bound together by tightly knit organised inter-firm transactions and collaborations that many academics and policy-makers seem to want them to be and that perhaps the image of cosy collaborations and friendly groups of scientists developing wonderful products after a short drive from home is fatally flawed. The evidence suggests that, for instance, these scientists might be better off driving to the nearest airport than the local business park [Bathelt et al. (2004)] and that if they are staying at home they are more likely to be innovative if they are enviously keeping watch on their competitors' achievements than if they are collaborating with them. On the basis of the evidence it seems that localised clusters are perhaps best understood as sites of informal social interaction and as arenas for flexible and well-functioning markets for specialised and skilled labour. In short, there seems to be little evidence that organised inter-firm transactions and co-operation characterises successful firms. At the same time, there is growing evidence that labour market dynamics and social interaction at the level of the individual can play important roles in firms' and clusters' knowledge creation processes.

Such remarks regarding the interplay between local and extra local (or global) collaborations (or lack of them) are a theme that is returned to later in the research when looking to find certain types of Web derived linkages and associated networks.

Chapter 3. Cluster Analysis Methods

3.1 Preamble

Cluster analysis is not a particularly new science. Long before it was applied to the study of the firm and industry, mathematicians, biologists, statisticians and a whole raft of other researchers had been looking at the way natural phenomena exhibited a tendency to 'group' together. Such clustering could be either by topography or by type or by some other attribute of the subject being studied.

It was therefore to be expected that such methods would be applied to discern the presence of industrial clusters given a firm population in a known geographical area. As a result of the availability of these tools regional scientists have used a range of methodologies for sorting industries into groups and which are discussed in outline in this chapter. However this study is not primarily concerned with a deep and wide evaluation of the various cluster concepts both new and developing. It is more about trying to find better ways of understanding the make up of industry in a particular locale that might be involved in a cluster of one or more of the types outlined thus far. In particular, as indicated in Section 1.3 the study is looking at ways of using the power of the internet to better inform the process of data acquisition and subsequent analysis in relation to research on industrial clusters. It is however necessary to have an understanding how such analyses are or have been carried out and this chapter gives an outline of the process.

There are a number of researchers who from time to time undertake a review of the state of knowledge for both theory and practice as far as industrial clusters are concerned and these have been cited as a contribution to the review in this section [Kuah (2002); Bergman and Feser (2000)]. The approach here is a summary of how the main types of analysis for determining the presence of clusters would be carried out in an industrial context. In addition and as noted previously the review is aimed at providing a backcloth sufficient to enable gaps in knowledge to be identified in the context of the research that is to be carried out.

3.2 Analytical Approaches

There are at least six basic analytical approaches, ordered approximately in terms of how often they have been used although this tends to vary with each country, driven in part by the data available to the researchers.

The approaches are:

- expert opinion
- location quotient
- trade based input-output analysis
- innovation input-output analysis
- network analysis
- surveys

It is perhaps worthwhile also having a further subsection under ‘miscellaneous’ on the basis that from time to time researchers come up with some method that clearly does not fit any of the above categories, an example being that of ‘connectedness’ amongst the members of an academic or research institute as noted previously in ‘Other forms of Clustering’ on page 38.

Expert Opinion

Probably the most widely used method of finding the ‘make-up’ of industry in an area is through the gathering of key information from significant actors and decision makers in the locale under study. Important sources of information can be gleaned through “*agents who know the region’s industries in terms of basic practice, supply chains, current investment patterns, export characteristics and potential opportunities for new products…*” [Stough, Stimpson and Roberts (1997), p.2]

Although this method can be time and cost effective given access to a competent set of experts it can be the opposite if the naïve researcher is faced with the task of attempting to glean information on a potential cluster using sources that may be lacking in understanding of the very concept of cluster studies and also with a much less than perfect knowledge of the main players and the interconnections between them. This method, although ostensibly straightforward can therefore be dangerously misleading through bias from key stakeholders or ‘experts in the field’ but who in practice may have limited knowledge of the fine detail of the various players in their sector or of their place in the economy as a whole.

Over-representation and location quotients

After expert opinion the next most widely used method of discerning evidence of clustering activity would seem to be the use of location quotients (LQ). This process involves the researcher looking, on a bounded geographical basis, at the representation of firms with nominated SICs or logical groups of SICs and comparing their density with that of a peer group in the wider firm population, usually that covered by a country. Any group of firms or even industries with a location quotient greater than 1 indicates that there is a greater than 'average' representation in the geographical area under study. This method therefore makes the not unreasonable assumption that such over-representation is a proxy for 'activity' which is an indicator that the presence of such activity is generated by the presence of an active industrial cluster.

An example of this is a UK wide study referred to as the Sainsbury Report [Miller, Botham, Martin and Moore (2001)]. The report was based on employment data and although supported by qualitative information, it would not be expected to identify all embryonic or aspirational clusters for example. The reason for this is that, the study to a large degree used Dun and Bradstreet (DNB) data (including employment) and which of itself uses SICs to identify types of firm activity. However at best the DNB data based is refreshed only every 2 years and SICs can take much longer to capture firms based on new technologies. This theme is returned to in Section 3.4.

What might seem a simple firm count however has to take recognition of a host of factors including the size of the firm with employment count usually being regarded as a robust measure of firm size.

In the LQ method linkages or association between firms and activity such as non-traded interdependencies are discerned largely by inference. For example if there is a strong electronics sector with a small average size of firms with a lot of new starts then it would be expected that a noticeable venture capital sector should appear.

The method also requires a fair assessment of the likely groupings that occur in practice, in other words there is significant art on the part of the investigator when it comes to putting together appropriate 'groups' or 'templates' of aggregated firms. [Feser and Bergman (1998, 2000)]

Of note, and consistent with Porter's view the authors in the above referenced study commented;

"We recognise that SIC codes do not capture the subtleties of industrial life and the complexities of cluster activity. The identification of these will be vital in further work to develop clusters."

Input – Output dependent methods

As noted in the preamble to this chapter cluster analysis is not a particularly new science and before it was applied to the study of the firm and industry, mathematicians, biologists and statisticians for example had been looking at the way natural phenomena exhibited a tendency to 'group' together. It was therefore to be expected that such methods would be applied to discern the presence of industrial clusters given a firm population in a known geographical area of which Brenner (2001) is an example of the genre. As a result of the availability of these tools regional scientists have used a range of methodologies for sorting industries into groups and have tended to move away from for example the discerning of clusters by surveys or location quotients. These methods tend to be more sophisticated calling upon the use of trade based data combined with statistical processing to discern appropriate groups. Such trade based analyses are referred to as input-output methods. But also include graph theory, triangularisation, multi-variate cluster and factor/principal component analyses. Czamanski and Ablas (1979) provide a review of early contributions. There are numerous other examples of input-output applications together with associated guides for use [Roberts, (1992), Abbott and Andrews (1990), Scott and Bergman, (1997), Hewings et al, (1998) and Roelandt and Den Hertog, (1999)].

Some countries, most notably the USA but including the UK to a limited degree, collect data on the trades between various firms in mostly manufacturing industries. Access to such data allows inspection of the strength of linkage between firms or groups of firms. As a result, input-output techniques have become one of the principal means of studying industrial interdependence. In the UK The Input-Output (I-O) framework breaks the economy down to display transactions of all goods and services between industries and final consumers in the UK for one year. Information is presented in two key products: Annual Input-Output Supply and Use Tables (I-O SUTs), and Input-Output Analytical Tables (I-O ATs).

Supply and Use Tables

The I-O SUTs show the whole economy by 123 industries (e.g. motor vehicles industry) and 123 products (e.g. sports goods). The tables show links between components of gross value added, industry inputs and outputs, product supply and demand. The I-O SUTs link different sectors of the economy (for example public corporations) together with detail of imports and exports of goods and services, government expenditure, household expenditure and capital investment. Producing I-O SUTs allows an examination of consistency and coherency of National Accounts components within a single detailed framework and by calculating Gross Value Added (GVA) for each industry group, sets the estimate of annual Gross Domestic Product (GDP). GVA measures the contribution to GDP made by an individual producer, industry or sector.

I-O Analytical Tables

I-O ATs are compiled from I-O SUTs data and other additional sources. These tables contain symmetric (product by product) tables, Leontief Inverse and other diagnostic analyses such as output multipliers. I-O ATs show separately the consumption of domestically produced and imported goods and services, providing a theoretical framework for further structural analysis of the economy, the composition and the effect of changes in final demand on the economy. The I-O ATs are normally produced every five years.

Importance of Input-Output work

Since 1992 ONS[9] has used the I-O process to set a single estimate of annual GDP. This is achieved by reconciling various sources of data used in compiling the income, production and expenditure measures of GDP. The I-O work also plays a central role feeding into many key ONS items such as chain-linking the production measure of GDP, Producer Price Indices, Regional Accounts and Environmental Accounts. In Scotland however Input-Output (IO) tables provide a more complete picture of the flows of products and services in the economy for a given year, illustrating the relationship between producers and consumers and the interdependencies of industries. The latest tables, published 30 November 2005, describe the economic activity in the Scottish Economy during the 2002 calendar year. These tables are also available for 2001, 2000, 1999 and 1998 tables, with earlier tables available upon request. The IO tables provide a wealth of detailed information about the purchases made by each sector of the economy in order to

[9] ONS – UK Govt. Office of National Statistics.

produce their own output, including purchases of imported commodities and their contribution to Gross Domestic Product (GDP). As with the more general UK tables the level of aggregation is relatively high at 128 industry groupings.

A particular problem with Input-Output data however (and indeed with many government agency derived data) is that of keeping figures up to date and hence relevant to fast changing industries. Even Scotland which seems to do reasonably well is 3 years behind in terms of yearly data and subsequent publication. Firms can only be surveyed on the basis of the previous year's figures but the job of cleansing and classifying the data is non-trivial and 2 years would seem to be the minimum possible delay even if significant resource was put to the task. What is clear is that for new industries, examples being nanotechnology, the many variants of internet driven business and new forms of specialist biotechnology, they simply do not appear on the radar of the IO tables due to a combination of the coarse graining of the levels of aggregation and also the fact that such industries are not captured easily by systems whose basic data may be 3 years out of date by the time is available for public analysis. A further problem is that the SICs used as the basis for much of the data collection are of themselves only refreshed infrequently such that it can take up to 10 years for a 'new' activity to acquire its own SIC code.

This latter problem is not just one of official data and statistics applied to IO Tables and indeed it is a key difficulty for many data collectors, both public and private alluded to later in the thesis starting with Section 3.3.

Innovation drivers

A variation on input-output derived trade linkages is that associated with the interaction of innovation rather than (or sometimes combined with) traditional production flow matrices. Debresson (1996) outlines a comprehensive source for techniques and examples of such analyses. This method links with 'The Innovation Environment' on page 33 taking as its basis the diffusion of innovations.

As noted by Roelandt and Den Hertog (1995, p.5), the principal advantage of innovation matrices is *"their focus on actual innovation interdependency and actual interaction between industry groups when innovating"*.

The principal disadvantages to this kind of approach are the costliness of data collection and conceptual difficulties in survey design to accurately define and track the diffusion of innovations within some supposed cluster of organisations.

Network Analysis

The use of network analysis for looking at relationships and connections between firms and other organisations with whom they have a relationship is a relatively novel and new way of study. It is a form of input-output analysis but with the emphasis on the network, the chain of interdependencies as much as the quantification of trade flows.

It can also be looked at qualitatively using techniques from social network analysis [Wasserman and Faust (1994)] being the standard text on the subject. Debresson (1996, pp167-173) outlines a number of techniques for identifying clusters by directed graph [see also Debresson and Hu (1999)].

A number of researchers have used these techniques to gain useful insights into the diffusion of innovations [Assimakopoulos, (2000)] and these methods could be extended into more general cluster studies.

There are a number of network mapping tools available such as Krackplot [Krackhard, Blythe and Mcgrath (1994)] and UCINET [Borgatti, Everett and Freeman (2005)]. In addition to multipurpose mapping tools there have been a small number of exercises using molecular modelling software to try and simplify the inherent complexity of interaction and linkages present in even a small cluster [Freeman, (1999)].

It would seem that most of the work has been done with a relatively small number of actors as the level of complexity increases markedly as the number of actors (nodes) in a network increases. This is a corollary of Metcalfe’s Law[10], which, dependent on the publication, says that the value of a network increases exponentially with the number of nodes. However of all the popular ideas of the Internet boom Metcalf’s ‘Law’ was one of the most dangerously influential and misleading as it tended to be used out of context and in inappropriate cases.

The law is said to be true for any type of communications network, whether it involves telephones, computers, or users of the World Wide Web. Whilst the notion of ‘value’ is inevitably somewhat vague, the idea is that a network is more valuable the more people that a person can call or write to or the more Web pages they can

[10] Named after Robert M. Metcalfe, one of the inventors of Ethernet. The law was named in 1993 by George Gilder, publisher of the influential Gilder Technology Report but it was at least a dozen years old at the time of promulgation.

link to. Whilst this may be true it is important to recognise that Metcalfe's Law is a rough empirical description, not an immutable physical law.

Metcalfe’s Law gained traction at the time of the first dotcom revolution and anything that seemed to be able to explain the massive increases in network size (and associated stock values) of many dotcom companies was seized upon without question.

In a more considered view and with the benefit of a least another 15 years of internet growth other authors put forward a proposition for ‘value’ in networks based on the formula V = n log (n) where V is the undefined parameter ‘value’ and n is the number of nodes in a network [Briscoe, Odlyzko and Tilly (2006)].

Even though this formula gives much more modest (and realistic) results for value with increasing ‘n’ than did Metcalfe’s Law, clearly for increasing node count there is a disproportionate increase in the value of the network

Similarly when dealing with the matrix manipulation of a medium sized network of say 1000 firms the matrix is going to be 1 x 10^6 cells in size. By comparison most pure social network analyses dealing with say personal relationships amongst individuals takes place typically with a population of a few tens of actors. Thus even if solid relationships can be determined between actors, be they firms, organisations or related or supporting activities and which of itself can be a major task as discussed later, a problem of processing may still exist. This theme is discussed at length in Chapter 5.

The clustered nature of the economy using networks has also been documented by Kogut and Walker (2001) who investigated firm ownership in Germany, mapping out the links between 500 non-financial corporations, 25 banks and 25 insurance firms. Their analysis of the obtained company network is discussed further in page 84 but it clearly indicated that German companies are part of a small world [Watts (1999)].

There is also an issue of the differences between clusters and networks and this is also discussed later in the study. Potentially, in the context of the research question, the study of networks could yield new knowledge in the same way that accessing web derived data can be shown to support identification of the firm’s activities.

Surveys

Straightforward surveys, as a tool for cluster investigation, when set against some of the more exotic forms of statistical analysis, might at first sight seem a little mundane. However surveys may be necessary to acquire data for some of the other methods discussed in this chapter. In principle one can survey regional firms and non-local trading patterns, co-operative alliances and so on. It does require careful sample design as a random set of firms might miss out some key players. Firm populations are not homogenous. It also requires some considerable co-operation from firms particularly when trying to elicit semi or indeed completely confidential information such as key supply chain firms, other trading partners, customers, the educational supply chain and so on. The task is eased if such a project is for example, initiated or sponsored by a single large firm [Sadler, (2004)] or a Trade Association and endorsed by its constituent members but can still suffer many of the drawbacks associated with expert opinion regarding bias and of course not all members of an appropriate cluster would necessarily be members of such an Association or suppliers to a large and influential firm. Whichever route is taken it does require a lot of manual effort which perhaps indicates why survey based cluster studies seem to be rare.

Other Methods

Other unique methods of tracking or discerning the presence of industrial clustering do appear from time to time such as looking at linkages derived from individuals e.g. the alumni of a particular institution in the manner of the tracing of the lineage of a family tree as noted previously in ‘Other forms of Clustering’.

Even at this stage a recurring theme with cluster analysis methods is that associated with the quality, timeliness and robustness of some of the fundamental data used to support subsequent analysis. This topic is taken up in Section 3.3.

3.3 Problems with data

Data on the firm exist in many forms available via a number of public and private agencies. Examples of public agencies in the UK are the Office of National Statistics (ONS) through the Inter Departmental Business Register (not generally available) and NOMIS which provides detailed labour market data.

Of the private agencies there are the main credit rating companies such as Dun and Bradstreet (DNB), Experian and Market Location together with a variety of trade directories, phone number derived data from BT and Yellow Pages all of whom sell firm data at varying levels of aggregation. There are in addition a number of specialised ‘data harvesters’ gleaning data from sources such as internet domain name registrars.

There are however some pitfalls in using such data for academic research for at least four reasons:

(i) Agencies sell data that has usually been collected for a purpose other than academic research. For example with a credit rating database the main purpose is to collect data that allows a creditworthiness score to be automatically calculated and which uses mostly financial data.

(ii) The issue of accuracy. For the largest agencies they have up to 3 million records on UK businesses and even with a large call centre checking details, some individual records can be up to two years out of date.

(iii) Credit rating databases may not directly capture those firms that do not apply to the agency to be rated and as 90% of firms employ less than 10 people in the UK a very substantial ‘hidden economy’ exists, at least as far as the credit rating agency is concerned. To alleviate this problem some agencies augment their data with input from other sources such as Companies House for new registrants but again small firms who are sole traders and partnerships are not so registered. In any case small firms are not required to file anything much more than rudimentary financial and employment data.

(iv) Errors in data supplied to the agency (either public or private) by the firm. Those persons completing the enquiry form may make mistakes or may simply not know the answer to queries regarding the firm’s primary and secondary activity in the context of the language of the SIC. In the matter of a firm deciding it’s own SIC, sometimes the very term is new to the individual charged with the task and they may go for the first approximation.

The net result is that researchers looking to find strong evidence of clustering by the use of fields such as SIC, employment, turnover etc. as a basis for subsequent analysis would, in the view of this author, be advised to find some additional and independent form of confirmation of data integrity.

3.4 Problems with SICs

In addition to the more general remarks concerning firm data there are particular problems concerning the use of SIC codes. Most cluster studies start with some sort of database of firms and other economic entities and try and discern patterns of common or supportive activity, of trade, of knowledge exchange, of networks of co-operation and if possible the presence untraded interdependencies although this latter group are often discerned more by inference than by direct observation.

By far the most popular starting point in putting together firm ‘commonalities’ is the use of SIC codes in order to establish the activity that an individual firm is engaged in. The two most common SICs systems in use in the UK have been the UK system [United Kingdom Standard Industrial Classification of Economic Activities (UKSIC)] and also those of the US, for example Dun and Bradstreet uses USSICs/NAICS in all countries. There are also a number of others such as the EU brokered NACE (Statistical Classification of Economic Activities) and the CPV (Common Procurement Vocabulary) although the latter tends to be used in the context of larger pan European Union purchasing contracts via the OJEU[11].

A Standard Industrial Classification (SIC) was first introduced into the United Kingdom in 1948 for use in classifying business establishments and other statistical units by the type of economic activity in which they are engaged. The classification provides a framework for the collection, tabulation, presentation and analysis of data and its use promotes uniformity. In addition, it can be used for administrative purposes and by non-government bodies as a convenient way of classifying industrial activities into a common structure. The system is now identical to the EUROSTAT System NACE at the four digit class level and the United Nations system ISIC at the two digit Divisional level.

Historically the impetus for the introduction of these codes was to enable government agencies to better understand the make up of industry when determining the national stock of businesses and their likely effect on the national economy.

With regard to the UK, at the time of writing, the latest implementation of UKSIC codes was 1st January 2003 www.statistics.gov.uk/methods_quality/sic

[11] OJEU is the Official Journal of the European Union (formerly known as OJEC). This is the publication in which all contracts from the public sector which are valued above a certain threshold must be published.

A new version of NACE also came into effect at the same time and UKSIC(2003) as noted above is based exactly on NACE Rev. 1.1 but where thought necessary a 5 digit system was added. NAICS – the North American Industry Classification System http://www.census.gov/epcd/www/naics.html has now replaced the USSIC system, the latest implementation at the time of writing being NAICS 2007 which includes revisions to NAICS 2002 across several sectors. The most significant revisions are in the Information Sector, particularly within the Telecommunications area.

In the case of Dun and Bradstreet (DNB) and other similar companies they are primarily credit rating agencies and data extracted from the firm is sold on to a variety of interested parties including researchers. As noted above regarding the acquisition of firm data generally it is always important to bear in mind that the source and impetus for acquiring such data in the first place was to support the credit rating side of the business rather than academic research into clusters or other similar endeavours. A full description of how DNB data fields are organised is found in http://www.zapdata.com with the detail in http://www.b2bsalesandmarketing.com/downloads/Q106datadictionary.pdf. Of interest is that DNB have one primary SIC (8 and 4 digits) and up to 6 secondary SICs with a line of business description provided by the business itself.

All of these SIC classification systems have their shortcomings. The main ones are:

(i) A level of detail that is too coarse to differentiate activities that are similar, but not the same. This was the case with the USSIC at the 4-digit level. This type of problem is addressed from time to time by the authorities expanding both the scope and the resolution of the numbering system with respect to the level of detail for industrial activities. This of itself can lead to the next problem as in (ii) below.

(ii) A level of detail that endeavours to cover every manufacturing and service activity known to civilised man but which may end up confusing the organisation asked to place itself in any particular category. Further, even quite small firms engage in more than one activity yet most questionnaires used as the basis for eliciting SIC data allow for one primary and a small number of secondary classifications.

(iii) The use of more generalised categories such as ‘Business Services’. In the event that it is not immediately obvious to the firm that a highly specific category exists for their particular firm's activities there is a temptation to place their organisation in this seemingly 'catch all' category. This is particularly true for service or knowledge based firms or those with a consultancy content across a wide range of services. With the rise of new types of firm activity such as those associated with e-business their development outstrips the capability of the authorities to keep pace with the emergence of such firms by the issuing of new SIC codes. Even with the best of intentions on the part of national authorities there is a natural time lag for an agreed new set of SICs to propagate through either the public or private industrial data collection systems. With classifications such as NACE there is the additional difficulty of getting all the EU member states to agree on changes for a common coding system although as noted above the UK system has been completely conformant since 2003.

The above shortcomings are well enough known by investigators and for some data sources such as Dun and Bradstreet (DNB) a text descriptor of the firms activities is often available but again this is applied to the main line of activity rather than an attempt to capture a full understanding of the subtleties of diversified activities that so many modern firms are engaged in. This can be a particular problem with the trend towards flexible specialisation in manufacturing firms [Murray (1987, Sabel (1989), Van Dijk (1995), Heidenreich (1996)]

It is therefore such compromises that can undermine the accuracy of analyses, no matter how technically sophisticated the subsequent statistical processing. This difficulty was articulated by Porter (1998, p204) when he complained, in relation to cluster boundaries, that they

"rarely conform to standard industry classification systems which fail to capture many important actors in competition as well as linkages across industries ….Because parts of a cluster often fall within different traditional industrial or service categories, significant clusters may be obscured or even go unrecognised".

In a further example [Feldman, Francis and Bercovitz, (2005), p. 131, 139] note

"Innovative firms often defy classification by standard schemes as they create an industry segment by responding to market opportunities typically operating in niches not profitable for larger or established firms" and later they conclude *"It is only*

through an appreciation of the nuances of cluster development that one might be able to inform policy adequately".

Similarly in a more recent paper by Henry, Pollard and Benneworth (2006, p.282) the same problem is still being articulated as part of a plea for a robust and multi-stranded cluster methodology:

"Initial identification of agglomerations use location quotients and input-output analysis to provide possible candidate clusters in any particular industry. More sophisticated technical variants may be applied in this analysis and, for example, might include aggregation of certain industrial classifications. If input-output relationships provide one justifiable rationale for aggregation, others might exist such as expert knowledge of emergent production systems (e.g. biotechnology) or systems 'hidden' within official classifications (e.g. the UK motorsport industry)"

Problems of Measurement

Although not strictly related to the problems identified with SICs it is also worth noting here that there is a fundamental difficulty regarding the measurement of what might or might not constitute an industrial cluster. In the literature and as noted above there are a range of techniques to measure and spatially delimit agglomerations of industrial activity. However it seems that basic empirical questions concerning what level of industrial agglomeration should be taken to indicate the possible existence of an industrial cluster remain largely unanswered in that there appears to neither consistency nor agreement regarding theoretically determined cut off points for defining such a critical level of agglomeration [O'Donoghue and Gleave (2004)]. These authors do put forward a case for a standardised location quotient method based upon aggregate data to identify those locations that have exceptional concentrations of activity as found on statistically significant basis. This is thus different from the more usual arbitrarily defined cut off bases in Location Quotients such as 1.25 in Miller *et al* (2001) or Maskell and Malmberg (2002) with an LQ larger than 3. The point being made here is that at the time of writing there is not universally agreed metric for what constitutes the measure of a cluster whether that cluster is by firm count, by employment or by network activity. This is an issue that is retuned to later when trying to compare a variety of agglomerations found as a result of the research

3.5 Summary of cluster considerations.

It was stated in the introduction that this was not to be comprehensive review of definitions or methods of discerning and measuring cluster activity amongst firms and other organisations. The intention is broadly to gain an understanding of the science (and art) sufficient to highlight some of the historical and current shortcomings and difficulties in cluster analysis which will inform subsequent research.

There does not seem to be any 'one way' to analyse clustering activity and indeed much depends on the desired scope of study on the part of the investigator. It would seem also that a combination of methods and data sources may yield the best understanding but the downside is that such an approach requires time, resources and effort on a number of fronts and hence is expensive.

To reiterate, Porter (1998) and Bergman and Feser *et al* (1998) have highlighted the fact that SICs whether US, EU or UK do have fundamental drawbacks in use and that there are often 'unseen' networks of activity not disclosed no matter how esoteric the statistical inferences subsequently used in computation. As an example there is a fair degree of activity in Defence related industries in the North East of England stemming in part from an historical connection in armaments manufacture and other industries such as naval shipbuilding and mechanical engineering with a military market bias. However an SIC based search for defence related activity reveals very few such firms. On the other hand a search for firms that could contribute shows many but based on SICs only there is no direct evidence that they actually do form part of a 'defence' related industrial agglomeration.

Another well known example, alluded to by Henry, Pollard and Benneworth (2006) above is that of the motorsport cluster [Henry and Pinch (1999) and Henry (2003)] which is made up of firms with SICs that the researchers felt should contribute to the cluster. The reason that their study was done this way was because at the time there was no SIC code for 'motorsport' existed. In a further similar example Raven and Pinch (2003) looking at the British kit car industry felt that there were some undiscovered parts of the industry that limitations of the SIC code system did not reveal.

It seems from investigation so far that in the matter of using conventional SIC based firm data that there are shortcomings serious enough to adversely affect final outcomes and hence the accuracy and any conclusion on the presence (or absence) of industrial clusters in any particular locale. It is the intention in the main part of the research to look at a new method of determining a comprehensive picture of firm activity free from some of the constraints of an SIC based system as outlined in this chapter. This will be carried out with reference to a defined geographic basis, in this case an English region. The next chapter therefore, whilst still being a continuation of the literature review, looks to try and find ways of improving the descriptors of a firm such that it gives an additional insight into what firms do, particularly in the case of a diversified firm serving a number of often seemingly unrelated markets.

The basic method will be to access a firm's web site and then interrogate that site for key words that could be descriptors of activity. This is a simple enough task for individual sites and when done manually. However, by definition, data will be required on enough firms in a geographical region to find related activity and it is not economical to visit thousands of URLs by manual look-up and interpretation.

In future chapters it will be shown how a method for mass access of known firm websites will be developed and deployed with a view to extracting key words and/or phrases from web pages in order to determine activity similarities and characteristics to be expected of industrial clustering phenomena. The results of themselves are unlikely to reveal clusters directly but if the technique can be demonstrated to contribute to knowledge over and above that from current methods, it could be part of a wider study in combination with some of the techniques outlined in preceding chapters.

What follows in the next chapter is a continuation of the literature survey, this part being focussed on the methods and tools necessary to develop such a programme.

Chapter 4. How can the web help in cluster research? – A few clues from the literature

4.1 Preamble

This chapter is part literature review and part methodology being written on the basis that some investigation is required in order to best guide the subsequent literature search process. The same philosophy applies to the following chapter on Networks.

The phenomenal rise of the internet and the World Wide Web over at least the last 15 years together with specialist search tools has opened up a variety of possibilities with regard to fast access of information and on a global scale. This leads to the tantalising prospect of acquiring new knowledge from existing web extracted information provided that the sheer scale and almost infinite variety of such information can be handled. Clearly methods of focussing down on the sheer scale of the internet are key. The page count of the internet climbs inexorably upwards; 8 billion pages [Search engine watch, (November 2004)], 11.5 billion indexable pages [Guilli and Signorini (January 2005)], 25 billion pages on Google in 2006 [http://en.wikipedia.org/wiki/Google (search_engine)] with the latter organisation having noted in a patent application in May 2006 that the estimated actual size of the web (as opposed to that which has been indexed) at 200 billion pages [http://www.seobythesea.com/?p=193].

The idea of using 'information' extracted from text sources including more latterly the web and then processing it automatically to discern patterns is not new. It has been used for a variety of tasks from finding market prospects to medical discovery to company activity on a country wide basis. With the rapid growth of digital information resources, information extraction (IE), the process of automatically extracting information from natural language texts, is becoming more important. A number of IE systems, particularly in the areas of news/fact retrieval and in domain-specific areas, such as in chemical and patent information retrieval, have been developed in the recent past [Chowdhury (1999) and Lawson et al (1996)] using the template mining approach that involves a natural language processing (NLP) technique to extract data directly from text if either the data and/or text surrounding the data form recognizable patterns. When text matches a template, the system extracts data according to the instructions associated with that template. A logical extension to these processes is to use words or phrases extracted from a web search or specialised crawler[12]. The combination of these two distinct processes is sometimes referred to as 'web mining' [Craven et al. (1998)].

4.2 The development of Tools.

Web mining investigation relates to several research communities such as Database, Information Retrieval, natural language processing and Artificial Intelligence.

Although there is some confusion about Web mining, the most recognised approach is to categorise Web mining into three areas:

- Web content mining,
- Web structure mining
- Web usage mining.

Web content mining focuses on the discovery/retrieval of useful information from Web acquired contents/data/documents, while Web structure mining is concerned more with the discovery of how to model the underlying link structures of the Web.

Web usage mining is a relatively independent, but not isolated, category, which mainly describes the techniques that discover the user's usage pattern and tries to predict the user's behaviours. This latter is perhaps of less interest in the context of eliciting information to enrich SIC descriptors, however it might be of some value in determining interactions by users and hence contribute to wider cluster study.

The challenge of having computers to not only gather and represent existing knowledge on the web, but also to use that knowledge for planning, acting and creating new knowledge can be thought of as a stepwise process. The first part of initially gathering knowledge and then mining it has been addressed by a number of Authors. The wrapper[13] induction community, for example [Kushmerick et al. (1997) and Knoblock et al. (1998)] has developed learning algorithms for extracting propositional knowledge from highly structured automatically generated web pages.

[12] **Web crawler** A piece of software (also called a spider) designed to follow hyperlinks to their completion and to return to previously visited Internet addresses.

[13] Information extractors from automatically generated text are usually called wrappers

The information extraction community [MUC-4 Proceedings, (1992)] is oriented more towards extracting propositional knowledge from free form, unstructured data sources.

The goal for these latter techniques is to reconstruct in symbolic form knowledge known by the original author and represented explicitly in the text of the web page in question. The field of study has progressed from hand constructed extraction rules [Soderland, Lehnert, (1994) and Soderland, Fischer and Lehnert, (1997)] to the 'learning' of extraction rules from a set of data. An example of the latter uses self learning rule-based information extractors to identify the name of a person given their home page [Frietag, (1999)].

Another slightly more sophisticated approach deals with extracting relational knowledge existing on the web through a combination of web pages and their hyperlink structure. The goal here is to look beyond the formatted text on the web pages and to learn to identify relations suggested by hyperlinks between pages. An example of this is given by Slattery and Craven (1998). This is potentially very interesting for any cluster study as it holds out the prospect of not only finding interesting and common key words - a form of 'keyword related location quotient' but also some relationship between activities within diversified firms.

These three types of information gathering techniques have been integrated by Craven, *et al.* (1998, 2000)

Most of the researchers and practitioners in the field seem to be part of a computer science or mathematical sciences community with a bias towards artificial intelligence and computational linguistics [Agrawal, Mamilla, Srikant, Toivonen, and Verkamo (1996)], [Ahonen, Hemonen, Kiemettine and Verkamo (1997)], [Barish, Knoblock, Chen, Minton, Philpot and Shahabi (1999)], [Hearst (1999)], [Mitchell (1997)], [Quinlan and Cameron-Jones (1993)], [Quinlan 1990)] and as such are interested particularly in the development of the technique. However the practical application of these techniques seems to be of interest largely to demonstrate the efficacy of any particular program in the harvesting of knowledge. Researchers tend therefore to pick applications that either support their research or test and demonstrate the scope of such research. As a result it would appear that the examples used to demonstrate the capability of any particular extraction and analysis algorithm in the field of industrial research in general and analysis at the company level in particular are few in number.

4.3 An Example in Context

One example however that does touch on the usefulness of such techniques applied to a population of firms, is an attempt to gain knowledge from the analysis of the Hoovers On-line Web resource, [Ghani, Jones, Mladenic, Nigam and Slattery (1999)]

In their paper the authors use Hoovers[14] Online Web resource as a starting point for discovering information and links from 4312 company websites. Hoovers Online is a web based resource of company information and which has, for the number of companies visited, a published URL. The methods used searched the first 50 pages on each site using a crawler to extract this information where possible. In all just over 108000 web pages were visited.

These pages formed the basic information input for subsequent processing but further data was also extracted from the wrapper, i.e. the standard Hoovers site rather than the company URLs. This data was added to the knowledge base. In addition to the extraction of text the features of the program concerning links were used to answer queries of the form ‘companies linked to by this company who link to this company’.

The subsequent research and analysis is beyond the scope of this paper, although part was undertaken using the commercially available text and data mining program ‘Clementine’ [SPSS]. The results achieved are able to derive knowledge of the form, for example;

‘Companies headquartered in Madrid having listed historical financial information use Arthur Andersen as their auditor’ and ‘Advertising agencies tend to be located in New York’. Whilst these may seem obscure outputs it does give some idea of the power of such analyses.

[14] Hoovers Online (www.hoovers.com) is a US based site which covers a wide population of the largest companies, predominantly in the US but is also represented in the rest of the world. Amongst general company financial and other data a proportion of corporations have a URL listed.

4.4 Links

The notion of 'links' has a number of definitions. As noted above with the Hoovers example, links means some part of the wrapper where a spider can expect to identify a reference to another firm or associate that is of 'interest' to the company being visited.

An alternative would be a hyperlink within the company's web site on the basis that if a company references another through a hyperlink it has done so intentionally.

The notion of an 'interest' can be quite broad in application but essentially means that company A feels that if it references say company B or associate C or establishment D and so on then it may be part of its own sphere of interest. For company web sites this could include hyperlinks, client lists, suppliers, supporters and sponsors and trade references.

This theme is explored further in Chapter 10.

4.5 Extracting text from the Web

One of the objectives of this thesis is to assess the practicality of extracting information from the web that can be used to enrich the more usual SIC based descriptors on individual firms. The preamble has indicated how such an objective might be approached and how other researchers are developing tools that might be usefully used or adapted in order to take this process forward.

The World Wide Web is a largely unregulated medium for information dissemination and as a consequence of this the data contained within individual web pages is usually unstructured and 'noisy'[15]. In general terms the starting point for the automated reading of web page derived text is to use some form of web crawler or

[15] The term 'noisy' used in the physical sciences sense usually refers to random errors in measurement or the presence of external and undesired inputs. In our case it is the latter or the presence of extraneous and irrelevant data and/or information that can cloud or obscure observation of the desired text or meaning. A treatise on the topic in relation to internet derived information can be found in [Pierre J. M. Data Mining from Noisy Learners. 2003, http://www.sukidog.com/jpierre/siam2003.pdf].

'spider'[16] that extracts and returns web pages against some specified set of instructions. For example a URL may be a given and the spider instructed to then scan individual pages for title, metatags and body text, to extract these and store them for further processing. The next stage would be to take such text and process it in same way that a human would do to extract keywords or to parse phrases or even concepts.

The Problem of 'noise'.

Even a simple excursion into the web using URLs for say a group of industrial companies shows that the above approach is fraught with problems due to the mass of 'messy' text and other data that characterises most web pages. Examples of such problematical causes in broad categories are:

- Copyright and other legal warnings
- Error messages of various kinds
- Unrelated download information such as Adobe Acrobat, Quicktime, RealPlayer.
- Non text information such as Flash sites
- Pop-up boxes and unrelated advertising
- Passwords and security issues
- Very large sites with multiple routes e.g. a University
- Duplicate text at several levels
- Corporate information unrelated to the activity of the firm i.e. stock prices, general news feeds.

The above relate to text that is seen by the reader. For any automated process some spiders may in addition pick up a great deal of additional programming instructions that are hidden to the normal human reader but which do exist on each page. The result of this is that some way has to be found to deal with this type of information that may obscure 'relevant' data or text in the search to find information that helps the objective of describing the firm and its activities.

[16] A spider can be thought of a web crawler with some inherent capability that allows it to act in the manner of an intelligent agent operating to a specific set of instructions.

Discerning Meaning from Raw Text

Assuming the noisy texts can be dealt with in some way we should be left with a collection of words that have within them useful descriptors or phrases describing the activities of the firm. A simple indexing system would find the most popular words for any record but this would not necessarily be useful. For example if the word ‘computer’ appeared many times against a record it could mean anything from a computer systems house to a call centre selling computers. It is exactly these kinds of subtleties of industrial activity that we are trying to differentiate. In this case it might be possible to use the original SIC as a ‘key’ or guide in some text processing program. However phrases such as ‘computer network design and implementation’ would be useful but a simple word parsing program does not ‘know’ which phrases would be of interest to the user.

Conclusion on the possibilities for data extraction and processing.

The World Wide Web has become a significant source of information. Most of this computer retrievable information is intended for consumption by humans and is not readily available as a data source in computer-understandable form. The current research challenge for this domain is to have computers not only gather and represent knowledge on the Web but also to use that knowledge for planning, acting and creating new knowledge. Some of the techniques being developed have a clear potential for a wide range of uses and driven by the research effort, in some cases funded through the military there is likely to be a consolidation of the most promising areas. It is also likely that other, hitherto underdeveloped probabilistic learning mechanisms will be further refined.

As previously mentioned there was an initial desire to use web derived data to augment standard SIC databases. Originally it was thought to manually look up a URL and then try to decide which words on the company’s web pages best described what is was the firm was about and the range of its activities. These words would then be transferred to the original database that the URL came from as a field along with other company data. This process would be fine for a small number of URLs but very time consuming for a large number. Even a cursory calculation indicates 70+ working days for the 4000+ URLs proposed initially and clearly much longer for say even a small region such as the North East of England with around 30000 significant firms and about the same number of other smaller but hard to find from a data point of view. To try and automate the process, a somewhat simplistic approach was assumed. This was that web pages would be extracted from a given URL and initially a simple keyword search would be implemented to gain an idea of the most popular activity. As diverse and unrelated sites tend to have many common words e.g. computer or Adobe Acrobat (as many sites have this link early on to enable download of subsequent .pdf files); a standard data mining package would be applied to ‘see what emerged’. Ghani et al. (1999) however showed that such an approach could be full of errors and in addition many opportunities to improve the accuracy of information extraction could be missed. Such an approach would also be likely to say little about links between firms. Although the above referenced authors did at one point use a standard data mining software package the pre processing carried out was a necessary and non-trivial part of the total analysis.

4.6 General Conclusions and the basis of a route forward

To access web based material without human intervention it is necessary to first gain access to individual web pages, look at the text contained within HTML tags either within the title page, as metadata or as body text. Not all of these text sources are present in all websites.

In the last few years a number of researchers have looked at ways to automatically extract meaning from the web. The research community for this activity has for the most part come from a computer science or machine learning or artificial intelligence background as noted previously. Early applications were used to spider the web on a topic specific basis and then to try to index or summarise the resulting large volumes of ‘harvested’ web pages. Other tools have been developed to take existing URLs and then to process the harvested text from individual sites using a combination of text mining and statistical analysis. In addition a number of commercial tools have appeared, often for marketing use and these can range in price from a few dollars to several hundred thousand. An example is a spider that searches out email addresses and goes from link to link to a specified depth. Such a tool comes free with the Windows 2000 and XP operating systems. It is however a relatively trivial task to identify an email address on a page compared to discerning meaning on a hands-off basis.

It is worth reiterating that examples dealing with data extraction related to the firm are few in number and those that have been undertaken have been done so by way of example to demonstrate the worth of some technique or process or to ‘see what emerged’ rather than to gain specific new knowledge about a particular industrial milieu

The literature review of cluster related activity highlights the diverse range of definitions, research and opinion regarding what is a cluster and how the presence and identification of a large number of variables have a bearing on the process. One of the key issues emerging is that of trying to improve the basic sources of data, both from an accuracy viewpoint and also with regard to the breadth and depth of the firm activity descriptors.

In many respects a literature search on topics related to the research here makes the tacit assumption that either existing research can be extended or existing methods can be applied in a different context to the original field of study or application. It is therefore necessary to make a judgement at some point as to whether the application of text mining in its various forms on web derived data could actually yield anything of true value in the context of exploring cluster activity amongst firms and other networked contributors.

Most available tools look for words and patterns in the web generally whereas we are proposing to investigate the reality of starting with a set of URLs, extracting words, phrases and information and then possibly applying some of the referenced text mining techniques to extract useful knowledge. Although there appears to be a fundamental difference there is a degree of commonality in the two approaches.

Obviously one legitimate outcome, although not a particularly satisfactory one, would be to show that, at this stage of the research the web cannot be relied upon to contribute anything additional and meaningful to the discerning of industrial clusters. At the opposite end of that continuum, ‘hidden’ or richer information about existing or ‘new’ clusters emerges which clearly indicates that the method works to some noticeable degree.

On balance, the literature would indicate that the objective of the proposed study can be fulfilled in some form although to what extent will only be found by true research and subsequent comparison with more traditional methods.

Chapter 5. Networks

5.1 Introduction

This section has been approached in a similar way to the first part of the literature review when considering the background to the study of industrial clusters in that sufficient historical and current theories and practice of industrial clusters were put forward in order to provide a backcloth for subsequent research in context. In some respects it could be argued that a literature review on networks should have preceded a similar review of clusters on the basis that clusters are a special case of or an extension of network effects. There is indeed scope for discussion to the degree to which a complex network with many actors is all that different from a cluster with ostensibly the same actors. This argument is acknowledged but the reason the reviews have been done in this order are twofold. Firstly in the examination of cluster theories, the majority of early authors use as their basis an economic argument in which the notion of networks, at least as far as our current understanding of the phenomena is concerned, are not initially considered much if at all. If we are following a chronology starting from say Marshall (1890) it is only a century later that consideration of networks starts to emerge with the work of economic geographers.

The second reason is related to the layout of this thesis in that Chapter 10 carries out research into linkages by the firm and takes as its starting point an understanding of the way networks function and whilst we bear in mind the economic ties that bind actors in an industrial cluster in some cases the links that join them look exactly like examples of pure networks albeit of different types as will be discussed in this chapter.

This section therefore covers the basic constructs of network theory and then later develops those constructs that have a particular relevance to the research question and as with the previous chapter it involves some preliminary investigation of method in order to best steer the literature review. As with text and data mining from the internet, the science and scholarship of networks is a formidable subject in its own right and this chapter seeks to lay out the basics that may inform future research. The fundamental link with previous chapters on the form and nature of industrial clustering and the internet is that network science is now known to be an integral driver of these two significant topics [Lichter and Friesz (2007)].

So far the term ‘networks’ has been used rather loosely and has been applied in a number of contexts, in particular Strategic Networks and Virtual Clustering in Chapter 2 and in the brief introduction to Network Analysis in Chapter 3. It is appropriate now that a more robust understanding of what we mean by Networks and their relevance to industry and the research question should be undertaken.

Early Work

The science of networks as a subject of study has been going on for about 250 years beginning mostly but not exclusively with Leonhard Euler in the 18th Century although his collected works were not published until sometime later [Euler (1913)]. Euler, the most prolific mathematician of his age introduced the notion of ‘graph theory’ by viewing networks and network problems as a collection of nodes and links. Euler further demonstrated that such graphs had *properties*, hidden in their construction that limit (or enhance) our ability to do things with them. After Euler’s death graph theory was studied by many of the mathematical giants of the time with the principal aim of discovering more about the properties of various graphs. Many of these problems related to ordered graphs such as the lattice in a crystal or the hexagonal lattice made by bees in a beehive. It was however two centuries before mathematicians moved from studying the properties of various graphs to asking the quintessential question as to how graphs, or more commonly, networks, actually came about in the first place.

The next major advance in the study of networks came about with the work of Paul Erdos and Alfred Renyi (1959 to 1997) who viewed graphs and the world they represented as essentially random[17]. To explain this idea: if we start with a large number of isolated nodes and we then randomly add links between the nodes in the manner of say random encounters between guests at a cocktail party. If only a few connections are added the only consequence will be that that some of the nodes will pair up. If however links are continually added then inevitably some of the pairs will

[17] A short treatise on Erdos-Renyi random graphs is given in http://www.math.cornell.edu/~durrett/math777/c1s1.pdf

connect to each other forming clusters of several nodes. When enough links are added such that each node has an average of one link a large cluster emerges. It can be observed that most nodes will be part of a single cluster such that, starting from any node we can get to any other by navigating along the links between nodes. Mathematicians refer to this phenomenon as the emergence of a giant component, one that includes a large fraction of all nodes. Different disciplines view the same phenomena from their own view and terminology in that a physicist would regard the emergence of the large component as a phase translation or percolation such as when water freezes and a sociologist would regard it as having formed a community.

The key message here is that when we randomly pick and connect pairs of nodes together in a network there is a point when something different happens and that is the network after having acquired a critical number of links drastically changes its structure. Where previously we had bunches of small isolated 'clumps' of nodes, disparate groups that only connected to or 'communicated' within a local cluster, the network is transformed such that there emerges a giant cluster, joined by almost all nodes.

The implications of such phenomena in the context of the literature on industrial clusters looked at thus far is that it may help to explain some of the successes of for example Silicon Valley and the Boston medical clusters [Chong Moon Lee *et al* (2000)].

5.2 Types of Networks

Small worlds and the (Six) Degrees of Separation.

In a highly innovative experiment, the psychologist Stanley Milgram (1967) sent a number of packets to 'sources' in Nebraska and Kansas, with instructions to deliver these packets to one of two specific 'target' persons in Massachusetts. The targets were named and described in terms of approximate location, profession, and demography, but the sources were only allowed to send the packets directly to someone they knew by first name. The object was to get the packets from source to target with as few of these 'first-name-basis' links as possible. Hence each link in the chain was required to think hard about which of their acquaintances would be most likely to know the target person or at least be 'closer' to them: demographically, geographically, personally, or professionally. Also, each link was supposed to record in the packet, details about themselves corresponding to those provided about the target, enabling the experimenters to track the progress of the packet and the demographic nature of chain along which it passed.

The conclusion from this experiment was that Milgram determined that a median of about five intermediaries was all that was required to get such a letter across the intervening expanses of geography and society. Whether this number is, in reality, too high or too low is a matter of debate and for example White (1970) proposed a revised model yielding an estimate of about seven intermediaries. The important point however is that whatever the precise number was, it was not very big compared with the overall magnitude of the system (the system in Milgram's case being the population of the United States which was about 200 million in 1967).

The above experiment has become well known because of its intriguing and perhaps counter intuitive nature but throughout the 1960's and 70's experiments by a number of other researchers [Granoveter (1973)], [Watts and Strogatz (1998)], [Barabasi A-L. (2003)] demonstrated that a variety of clustering and 'small world' effects could be found and quantified in all highly connected functioning networks, effects not explainable by random graph theory. These effects have been since found in a wide spectrum of real world networks including social networks, neural paths, the World Wide Web, the physical connections of the Internet, ownership of companies, food webs and cell biology. These studies demonstrated that extremely large networks do not need lots of random links; just a few long range links between clusters to enable small world features. Significant work done by Barabasi's team and in particular by Albert and Jeong (1999) indicated that for the World Wide Web any document is on average only nineteen clicks away from any other.

A final word on this – the six/nineteen degrees phrase can be deeply misleading because it suggests that things are easy to find in a small world. In practice this is far from the case. Not only is the desired person or document six/nineteen links away but so are all people or documents. In other words, six or nineteen can either be a very small number or a very large one, depending on what we are trying to do. Since the average number of links on any given Web document is around seven, this means that while we can follow only seven links from the first page, there are 49 documents two clicks away, 343 three clicks away, and so on. By the time we reach the nodes that are exactly nineteen degrees away, in principle we would have checked 1016 x 10^{16} documents, 5 million times more than the total number of pages on the Web. This contradiction has an easy resolution: some of the links we meet along the road will point back to pages that we have seen before. Thus they are not ‘new’ links. Even if it takes only one second to check a document, it would still take over 130 million years to get to all documents that are nineteen clicks away. Nevertheless, despite the abundance of choices, we sometimes find documents rather quickly, even without search engines.

This is, of course, because we do not follow all links but rather we use clues. Indeed, if we are looking for information on say ‘Wensleydale’ and are faced with four or more choices on a given webpage, we are more apt to follow the link to the locale than either the link for a cheese or a railway or a reference to the cartoon film featuring ‘Wallace and Gromit’ where Wensleydale cheese occurs as a key part of the script. However by interpreting the links we see, we immediately avoid having to check all the pages within nineteen degrees and can zero in on the desired page. Whilst this method seems to be the most efficient, it almost always fails to find the shortest path. Indeed, it is always possible that the Wensleydale cheese Webpage that we bypassed might augment their page with a link to a Wensleydale tourism site (as indeed it does). But most people looking for Wensleydale for say a holiday would ignore the cheese link and eventually follow a longer path to the destination. The computer, having no taste or bias, will chew through with equal efficiency (or inefficiency dependent on the viewer’s perspective) the location, cheese, railway and cartoon sites, pragmatically following the links to all of them. By trying all the possible paths, it will inevitably locate the shortest one, independent of the content of the intermediate pages.

Finding the desired page on the Web highlights a fundamental problem with six degrees: Milgram's method overestimated the shortest distance between two people in the United States. Six degrees is really an upper limit. There are an enormous number of paths with widely different lengths between any two people. Milgram's subjects were never aware of the shortest path to their target. This is like being lost in a huge maze where we can see only the corridors and doors next to us. Even if we have a compass and we know that the exit is toward the north, finding it could be woefully inefficient and time-consuming. With a map of the maze in hand, we could be out in a few minutes. Similarly, Milgram's letters would have followed the shortest path between Omaha and Boston only if all participants had had a map that compiled the social links of all Americans. Lacking such a map, they forwarded the message to those that they thought were most likely to take it in the right direction.

For example, if you wish to be introduced to the Prime Minister of the United Kingdom, you would try to think of somebody who knows the Premier. Most likely you would settle on your MP but as most people do not know their MP on a first name basis, we would try to find somebody who does and who would be willing to broker a meeting with the PM. That would take at least three handshakes. In the meantime, you might have no clue that the gentleman you sat next to a few days earlier at a dinner party went to school with the PM. Thus in reality you are only two degrees away from the PM. Similarly, the paths recorded by Milgram's experiment were invariably longer than the shortest possible. Thus the real separation in society was clearly overestimated. It must be shorter than six. We do not have a comprehensive social search engine, so we may never know the real number with any certainty.

Hubs and Connectors

In developing the chronology of network research it was found that ‘connectors’ have a pivotal role in networks [Gladwell (2000)]. In practical terms this can show in the sudden and dramatic spread of viruses, the success of fashion breakthroughs and other events characterised by a sudden transition to explosive spread. Throughout the world of scale free networks, those connectors are found as hubs. Studies of the World Wide Web for example [Barabasi (2003)] have shown it to be dominated by a very few highly connected nodes, hub Websites with many links that hold together large clusters and many not so popular and seldom noticed nodes.

Even though such hubs cannot be explained by the models used by Erdos/Renyi and Watts/Strogatz, they turn out to be ubiquitous in the most complex networks studied to date. The notion of the hub is not an artificial construct and in the context of firms and the links between them the importance of one or more key nodes acting as hubs will be shown to be highly relevant in the research phase.

The Emergence of Scale Free Networks

The economist Vilfredo Pareto observed the principle commonly known as the '80/20 Rule’ described mathematically by a ‘power law' distribution in a Chart or histogram of values. Unlike the familiar ‘bell curve’ characteristic of charts of random events, the power law distribution has no peak but is a curve in which a few large events coexist with many small events. This principle has generated numerous articles in the business world even getting its own book [Koch (1998)]

The number of links in a small fraction of the nodes is 'off the scale' of the large majority of nodes, such networks being referred to as 'scale-free'. In 1999, Barabasi's team found that these power law curves are consistently found in numerous large networks and he asserts that *"power laws are at the heart of some of the most stunning conceptual advances in the second half of the twentieth century, emerging in fields like chaos, fractals and phase transitions".*

Spotting them in networks signalled unsuspected links to other natural phenomena and placed networks at the forefront of our understanding of complex systems in general." [Barabasi, (2003 pp. 72-73)].

Preferential Attachment and the Notion of 'Fitness'

Studies of the evolution of system structures showed that hubs result from the combination of system growth and 'preferential attachment' to existing nodes in a scale-free model described earlier. Improvements in the model sought to include other phenomenon such as the appearance of internal links between established nodes, node disappearance and the rewiring of links due to ageing or retirement of nodes.

The explosive success of Google however caused something of a rethink in that Google's success in capturing so many links could not be explained by the scale-free model until the idea of node 'fitness' was introduced. This is a measure of a node's ability to stay in front of the competition. Examining the fitness model data, Bianconi (2001) found that the calculations used were very similar to those found in the formation of a Bose-Einstein condensate. The mathematics describing the behaviour of 'Bose gases' (a unique creature of sub-atomic quantum mechanics) turned out to be identical to those in the network fitness model. This similarity means, according to Barabasi, that in certain circumstances, particularly fit nodes in a network did not merely get 'richer' (in terms of network influence) but in certain circumstances a particularly 'fit' node could assume a winner-takes-all position.

In an ordinary 'fit-get-rich' network, the fittest node becomes biggest, but other fit nodes are close behind, so that "*the power laws and the fight for links are not antagonistic but can coexist peacefully*". In a 'winner-takes-all' system, the fittest node grabs all the links, shaping the network into a 'star' or 'hub and spoke' topology which is not scale free with the ultimate result that there is a single hub and many tiny nodes. These findings have obvious relevancy to those studying for example antitrust law and policy and the ongoing case of Microsoft.

The above systems however are not without their difficulties. Natural systems exhibit robustness, the ability to survive under a wide range of conditions, because of their interconnectivity and scale-free topology. Experiments on a model of the Internet showed that one could remove the majority, even 80%, of the nodes and the rest would hold together as long as the small minority of hubs survived. Even though they are robust in the face of failure due to random events or errors, scale-free networks do have an Achilles heel: they are vulnerable to simultaneous targeted attacks on the largest hubs.

With reference to the 1996 Western power blackout, Watts (2005) published a study of fads and ‘cascading failures’ in systems such as electrical power grids. Cascading failures are not unique to electrical networks. They have been observed in many complex systems, including the economy (for example the East Asian monetary crisis of 1997), in ecological systems, in cellular biological systems and following the September 11 terror attacks. Watt’s investigations resulted in a model he used to study such phenomenon and found that such cascades do not occur instantaneously; failures may go unnoticed for a long time before starting a landslide. Such phenomena invite study from many disciplines.

On the topic of so called ‘fads’ the spread of viruses and fads are examples of what is called diffusion in a complex network, with a calculable ‘spreading rate’. They will die out unless that rate exceeds a critical ‘epidemic threshold'. Random Graph Theory could not explain the persistence of certain computer viruses or the explosive spread of AIDS. When a scale-free model including highly linked hubs was used however, researchers found that the epidemic threshold vanished.

At this point it is perhaps worthwhile having something of a reality check regarding the place in the overall thesis of network theory considered thus far. Whilst there is a clear connection with the structure of the internet (and which is discussed further below) the relevance to industry in general and industrial clusters in particular may not be so readily apparent. In some respects however at this point network theory is still more of a backcloth for what is to come rather than a clear fit when applied to the way firms and other organisations interact with each other and with a regional economy. Nevertheless we can start to see relevancies emerging. Large groups of connected firms and other organisations exhibit traits observable in other networks far removed from industry such as biological organisms and as such the network has properties and whose behaviour may be predictable. Examples are the tendency of some nodes to assume dominant or 'key player' positions, a property clearly of importance in industrial cluster studies.

The next part of this review therefore attempts to close in on some of the wider network constructs particularly in an economic context and also with application to the Web.

The Rise of the Internet and the World Wide Web.

The Internet (on which the World Wide Web runs) has grown more like an evolving ecosystem than a manufactured machine. In 1999, three Greek computer scientist brothers [Faloutsos M, Faloutsos P, Faloutsos C. (1999)] found that the connectivity distribution of Internet routers followed a power law, demonstrating it to be a scale-free network and subject to the same vulnerabilities and robustness found in other such networks.

Unlike the Internet infrastructure however, the Web is a 'directed' network; its many links each point only one way, shaping its topology into the form of a 'bow tie' with four 'continents,' one being an archipelago of disconnected islands [Broder *et al.* (2000)]. It turns out that these same four continents are found in all 'directed' networks and are predictable analytically. This topology and its consequences for the navigability of the Web become relevant in legal and regulatory actions, such as the French court's order in the year 2000 that Yahoo must block French residents from navigating to neo-Nazi websites. Despite Lessig's (1999) assertions that code can enforce such legal directives, Barabasi (2003) maintains that the topology and navigability of the Web is a function of collective human actions using the code. The architecture of a scale-free, directed network represents a higher level of organisation than the underlying code. As long as individuals decide to what nodes to link to, the inherent topology (and navigability) survives according to Barabasi.

In other examples Dodge and Kitchen in their book 'Mapping Cyberspace' (2001) show the staggering topological complexity of internet mapping projects with upwards of 100000 main nodes, a figure that was exceeded almost the day after the computation was finished. Although this type of analysis may not be immediately be seen as relevant in the study of industrial networks and clusters, if however the nodes are replaced by industrial entities and if enough information on them can be assembled then all the underlying theories on networks become useful for both analysing and predicting cluster behaviour. Such is the thrust of the research from Chapter 10 onwards.

The Network Economy.

The traditional tree-shaped corporate structure, suited to mass production, is poorly suited to deal with rapid innovation and market change. The challenge of competing in such environments led to industries such as pharmaceuticals to adopt technologies based on scale-free networks of alliances and outsourcers. For years, economists spoke of a standard formal model of economics, in which companies interact not with each other but with 'the market,' a theoretical entity mediating economic transactions. Barabasi: "*In reality, the market is nothing but a directed network. The weight of the links captures the value of the transaction, and the direction points from the provider to the receiver. The structure and evolution of this weighted and directed network determine the outcome of all macro economic processes."* [Barabasi (2003 pp.208-209)].

Also in Powell (1990) *"in markets the standard strategy is to drive the hardest possible bargain in the immediate exchange. In networks, the preferred option is often creating indebtedness and reliance over the long haul".* As a scale-free network, the economy is subject to the same vulnerabilities as are power grids and the Internet. Failures can, as previously noted, cascade through the whole economy, as did the 1997 financial crisis that began in Thailand and cascaded across the Pacific, resulting in the stock market crash of October 27, 1997. Barabasi asserts that this is a natural consequence of network interconnectedness and interdependency and notes economic and network research in this field referencing many such studies, particularly those of Leigh Tesfatsion (1996 – 2006).

Besides being scale free, a network economy displays clustering as well. There is first a strong geography based form of clustering, where companies have more links to local consumers. Such an assertion however does assume that local consumers want what the company has to offer and by definition clusters must have very strong exports if they are to maintain above average concentration in a region.

The work of and Kogut and Walker (2001) was noted briefly in Network Analysis and is expanded here. Their study of the clustered nature of the economy using network methods investigated firm ownership in Germany, mapping out the links between 500 non-financial corporations, 25 banks and 25 insurance firms. In their network two firms were connected if they had a common owner. In some respects the network that was found was similar to an actor network, where the actors corresponded to companies and movies to owners. An example of this is the so called Kevin Bacon game[18] [Fass, Ginelli and Turtle (1996)]. A typical owner spans multiple companies, just as there are many actors in a single movie. The analysis of the obtained company network clearly indicated that German companies are part of a small world [Watts (1999)].

Interestingly the diameter of the network is 4.81[19] i.e. the majority of these companies are linked through a chain of four or 5 owners. Kogut and Walker found a huge clustering co-efficient as well. If the companies were to form a random network, the chance of finding a link between two neighbours of a certain firm was estimated to be 0.5%. In contrast in the real network two neighbours of any firm have a 67% chance of having a common owner. This is clearly a significant difference underlying the very high degree of clustering characterising the economy, at least as far as ownership connections are concerned.

It is of course quite possible for firms with such highly connected ownership links to have few trading links. For example if the firms were connected by venture capital owners, unless such owners were following a deliberate policy of investing in firms where it was possible to take advantage of buying and supplying links between the owned firms then evidence of clustering based on trading relationships would not necessarily emerge.

Connections via owners of companies are not the only way of studying linkages and there are many other studies undertaken by tracking the personal links of key players. Eisenhardt and Schoonhoven (1990) showed in their study on the formation of the US semiconductor firms that the personal network of management teams plays a vital role in assessing early network partners. Similarly Hite and Hesterley (2001) put forward the thesis that networks of individual firms consist primarily of socially embedded ties of the management teams as organisational networks emerge. The

[18] The ‘Oracle of Bacon at Virginia’ can be found at http://www.cs.virginia.edu/oracle/

[19] The term ‘diameter’ relates to the number of degrees of separation discussed in 5.2 Types of Networks

authors characterise these networks as being identity based. As the organisational network moves into the early growth stage, the member firms' networks evolve toward more ties based on the calculation of economic costs and benefits. The persistent ties between the core set of network partners are based on trust from frequent economic exchanges and a link in the evolutionary process from networks to clusters is seen (at least in this case).

A practical example would be the Director networks of top American firms. Because of the important role boards play in shaping the landscape of American corporate life, the web of directors has often been scrutinised in the business literature. However, only recently with the advent of methods to analyse complex networks have we started to understand to what degree the power of this web is rooted in its interlocked topology. In the director network each node is a board member linked to directors serving on the same board. With thousands of companies, each with about a dozen or so directors, this is a rather large web. In a study of this type [Davis, Yoo and Baker (2001)] looked at the most influential component of this web, focusing on the network of Fortune 1000 companies, made up of 10,100 directorships held by 7,682 directors. If each director were to serve on one board only, the network would be broken into tiny, fully connected circles, each the size of a single board. This is not the case however. Whilst 79% of directors serve on only one board, 14% serve on two, and about 7% serve on three or more. The measurements indicated that these few overlapping directors create a small-world network with five degrees of separation. Indeed, the distance between any two directors belonging to the major cluster, which contains 6,724 directors, was 4.6 handshakes on average. The small-world nature of the director web is due to the 21% of directors who serve on more than one board, since they are the ones who hold this complex network together. This study does not say much directly about industrial clusters, rather it points out that there are other ways that firms link which are not based on cluster knowledge or place-based competitive advantage. What is of relevance is that the relatively small number of overlapping directors have a disproportionate influence on the network in the same way that Granovetter (1973) talked about the strength of weak ties. It may be that when observing other types of links between groups of firms in a trading network the same basic topography applies but the key connecting directors are replaced by key connected firms.

To try and summarise the basic message from this preliminary look at the evolving science of networks Kelly (1995) with a degree of foresight remarked *"The Atom is the past. The symbol of science for the next century is the dynamical Net . . . Whereas the Atom represents clean simplicity; the Net channels the messy power of complexity. The only organisation capable of non-prejudiced growth or unguided learning is a network. All other topologies limit what can happen. A network swarm is all edges and therefore open ended any way you come at it. Indeed, the network is the least structured organisation that can be said to have any structure at all. In fact a plurality of truly divergent components can only remain coherent in a network. No other arrangement - chain, pyramid, tree, circle or hub can contain true diversity working as a whole*."

Although physicists and mathematicians may take exception to some of these statements, the basic message is an interesting one: there is a convergence going on between the evolutionary topology of living matter, the open-ended nature of an increasingly complex society, and the interactive logic of new information technologies.

A final word from Barabasi: "*Network thinking is poised to invade all domains of human activity and most fields of human inquiry. It is more than another useful perspective or tool. Networks are by their very nature the fabric of most complex systems and nodes and links deeply infuse all strategies aimed at approaching our interlocked universe*".

5.3 Network thinking applied to the firm

The discussion on networks thus far has outlined, in a purposely wide ranging survey, the historical and current thoughts of a number of researchers and practitioners. Although some of the examples given of different types of networks might be regarded as somewhat obtuse with respect to this study which involves industrial and other organisations it has been shown that even highly diverse and seemingly unrelated applications of networks often have embedded within them features common to all. So although part of this chapter is concerned with network theory it is an important precursor to understanding how the science of networks may be applied to the firm.

Networks of co-operation have always existed whether as simple artisan networks or more sophisticated forms leading to true Porterian ideas of clustering but what has really supercharged the whole process has been the advent of the information technology revolution. This has led to a whole new techno-economic paradigm emerging in the last twenty years or so with the overlay of the internet in not much more than the last decade.

As Freeman (2001) writes: *"A techno-economic paradigm is a cluster of interrelated technical, organisational, and managerial innovations whose advantages are to be found not only in a new range of products and systems, but most of all in the dynamics of the relative cost structure of all possible inputs to production. In each new paradigm a particular input or set of inputs may be described as the 'key factor' in that paradigm characterised by falling relative costs and universal availability. The contemporary change of paradigm may be seen as a shift from a technology based primarily on cheap inputs of energy to one predominantly based on cheap inputs of information derived from advances in microelectronic and telecommunications technology."*

Ernst (1994) has also shown that convergence between organisational requirements and technological change has established networking as the fundamental form of competition in the new global economy. Barriers to entry in the most advanced industries, such as electronics or automobiles, have sky-rocketed, making it extremely difficult for new competitors to enter the market by themselves and even hampering large corporations' ability to open up new product lines or to innovate their own processes in accordance with the pace of technological change. Thus, cooperation and networking offer the main possibility of sharing costs and risks as well as keeping up with constantly renewed information. Yet networks also act as gatekeepers. Inside the networks, new possibilities are relentlessly created. Outside the networks, survival is increasingly difficult. Under the conditions of fast technological change, networks, not firms have become the actual operating unit. In other words, through the interaction between organisational crisis and change and new information technologies a new organisational form has emerged as characteristic of the informational, global economy: ***the network enterprise***.

A definition of the networked enterprise that Castells (2000) puts forward in this context is:

"that specific form of enterprise whose system of means is constituted by the intersection of segments of autonomous systems of goals".

Thus, the components of the network are both autonomous and dependent vis-à-vis the network, and may be a part of other networks, and therefore of other systems of means aimed at other goals. The performance of a given network will then depend on two fundamental attributes of the network: its connectedness, that is, its structural ability to facilitate noise-free communication between its components; and its consistency, that is, the extent to which there is a sharing of interests between the network's goals and the goals of its component members.

In some respects this is a more modern and to a degree internet based form of earlier work done by Kay (1993) whose ‘Architecture’ was seen as one of the determinants of firm competitiveness. Architecture referred to the network of relational contracts with and around the firm. Firms may establish their relationships with and among their employees (internal architecture) and with their suppliers and customers (external architecture) or amongst a group of firms engaged in related activities (networks). This work again was an extension of work done by Porter (1998)

Much of this activity is by definition, related to multinational companies with a global reach. However besides these large global players’ small and medium firms in many countries - with the US (e.g. Silicon Valley), Hong Kong, Taiwan, and northern Italy hosting the most prominent examples - have formed cooperative networks, enabling themselves to be competitive in the globalised production system. These networks have connected with multinational corporations, becoming reciprocal subcontractors. Most often, networks of small/medium businesses become subcontractors of one or several large corporations but there are also frequent cases of these networks setting up agreements with multinational companies to obtain market access, technology, management skills or brand name. Many of these networks of small and medium businesses are transnational themselves through agreements that operate across borders as exemplified by the Taiwanese and Israeli computer industries extending their networks to Silicon Valley and, as noted earlier, the cultural ties of Indian origin technological entrepreneurs locating in Singapore.

Furthermore, multinational corporations have increasingly decentralised internal networks, organised in semi-autonomous units according to countries, markets, processes, and products. Each one of these units links up with other semi-autonomous units of other multinationals in the form of ad hoc strategic alliances and each one of these alliances (in fact, networks) is a node of ancillary networks of small and medium sized firms. These networks of production have a transnational geography which is not undifferentiated: each productive function finds the proper location (in terms of resources, cost, quality, and market access) and/or links up with a new firm in the network which happens to be in the proper location.

Steiner (2005) also has a view on firm networks through the medium of business webs in that *"business webs are groups of companies that collaborate on the basis of technological and economic standards to provide a product system"*

It is also noted that although considerable work on entrepreneurship and the growth of a single firm already exists, there has been less activity on exploring the establishment and growth of firm networks although some work touching on this aspect has been carried out. [Shan (1990); Autio and Garnsey (1997); Gordon and McCann (2000); Sexton and Landstrom (1999); Miller and Garnsey (2000); Murtha, Lenway and Hart (2001); Staber (2001)] have each looked at some aspect of firm networks although usually as part of a supporting construct for other studies. More recently Casanueva and González (2004) have looked at the interaction of social and information relations in networks of SMEs. These authors concluded that in general, applying the methodology and assumptions of social network analysis for the study of interorganisational relations would allow various existing lines of research to advance and help create new ones.

5.4 Summary on Networks

Networks have for the most part been treated by those who have studied the science as a construct that can explain many situations in the organisation of nature or indeed of civilisation generally. Although originally rooted firmly in mathematical theory, in recent years the application of network theory to what might be termed the real world has gained currency and there are those who see networks as being a controlling force for many key aspects of life involving a high degree of practical utility from the structure of biological entities to the behaviour of the stock market.

In the context of this research it would appear that network theory applied to the study of industrial agglomerations has much to offer and indeed the research question specifically seeks to include the role that might be played by networks. To summarise what role this might be and to provide an input into later work using the internet as the basis for finding connections between entities within an industry based agglomeration we can say the following:

- Firstly, all networks even seemingly random ones have an underlying mathematical construct that imbues the network with properties and enables it to be measured using some appropriate metric.
- Secondly not all networks have the same properties, such properties being dependent on the boundary conditions caused by the external environment in which the network operates and the relationship between nodes internal to the network.

To try and focus on what might be useful in the context of the research question; the emergence of scale free networks, hubs and connectors and the notion of ‘fitness’ or preferential attachment are clearly of interest because these elements can be seen to be at work within groups of interconnected firms. Similarly the literature on the network economy gives many examples of interconnectedness amongst firms and in addition there is often strong geography based evidence of clustering where the term clustering is here an attribute of a network.

The final part of this section was given over to the theme of networks applied to the firm and the part played by the advent of the information technology revolution. This manifestation of more generalist mathematical concepts of networks combined with a practical examples of the way modern firms interact with each other and with their external environment is a theme that will be returned to in later chapters when looking at the part that may be played by the internet in trying to discern the types of networks, the nodes, authorities and linkages that have been indicated here.

If a word of criticism is due it probably relates to the tendency of promoters of the network concept who see networks everywhere. In the context of the research question therefore it is important to keep the network concept rooted in practical reality so that what is a genuine network at work in an observed industry agglomeration is both evident and provable.

Chapter 6. General Conclusions on Literature Survey

This literature survey has covered three main topics that are deemed to be of interest for the subsequent research phase.

These topics are broadly:

- Cluster theory and practice
- Web structure
- Network theory

Each of these subject areas on their own have a significant number of papers, books, articles, web references and other citations but the literature search, as noted at the outset, has been selective and focussed on supporting the research question.

Whilst some of the underpinning mathematical and economic precepts for cluster studies and for networks can be traced back to the 18th century, for the most part the majority of studies have taken place within the last 25 years and for those related to the web within the last 15 years or so. As such the literature is still evolving and although the production rate of publications for topics related to the phrase ‘industrial clusters’ may have slowed in recent years the research activity on networks in general and in particular on web topics seems to be increasing, driven in part by the phenomenal rise in use of the internet and in part by opportunities for new web related products and services.

This latter phenomenon bears on some early remarks made in Section 1.3 where it was assumed that during the course of the research some tools and capabilities would be made available in Internet use that were not envisaged at the outset. This leads to a secondary problem in that it is necessary to constantly revisit the literature study as relevant papers regarding such web developments are being added at a significant rate.

To try and summarise the current state of thinking in these topics noted above:

<u>‘Traditional’ clusters studies</u>

The section started with a short review of some supporting theories particularly from a strategic management point of view on the basis that how value is created is an important underlying concept with regard to agglomeration of industrial companies and other organisations.

The review then moved on to an examination of forms of industrial clusters. These forms might be regarded as 'traditional' in the sense that their underlying ideas predate the internet era or indeed that of mass access to information technology. The influence of Michael Porter was then mentioned separately in this review but analysis deferred as Porter is discussed at length later. The review thus far was on the basis of scene setting in that it is not necessary to be too detailed in the analysis of historical concepts as the research question is more concerned with the influence of the internet and the power of latter day information technology. As such due attention was given to the influence of the internet and the use of virtual clusters, the latter being deemed to be free of spatial constraints thus obviating many of the hitherto carefully wrought definitions of clusters that saw geographical proximity as a fundamental requirement.

A final part of the preliminary review of cluster definitions was concerned with some of the challenges to received wisdom and this regard Porter has come in for particular criticism for the very definitional incompleteness of his cluster concepts and a general fuzziness over the balance of proof that clusters actually do any good in the sense of economic worth.

A point that was picked up even at this early stage was that commentators were voicing concerns regarding the lack of richness in data describing the firm which was used as the basis for subsequent processing.

The review then moved on to Chapter 3 whereby initially six defined methods used by researchers for discerning industrial clusters were outlined. It was noted that in most of the analytical forms of cluster analysis (as opposed to surveys and expert opinion) a database of companies and other relevant organisations was a pre-requisite. A review of the kinds of databases used typically by researchers was given with a critique of some of the known shortcomings of the structure of these databases, the accuracy of the data and the part played by Standard Industry Classifications (SICs). A number of prominent authors voiced particular criticism in this respect having found the shortcomings of the various national SIC systems not ideally suited for discovery of innovative firms or the newer types of organisation based on the exploitation of new technologies. They also pointed out that cluster boundaries rarely conform to SIC systems which fail to capture many important actors in competition as well as linkages across industries. SIC based studies also have difficulty picking up ‘factor conditions’ [(Porter (1990)] and associated supporting industries and the establishment of ‘templates’ to capture these are only partially successful.

These shortcomings when using SIC based data on the firm are key to the research question and the research that follows is, in part, an attempt to overcome some of the known drawbacks of SICs articulated in this section on cluster methods.

Web Structure

The research direction regarding the literature on web topics was initially based on the idea that the internet, as a vast information resource, must be able to add something to the identification of firms, industry and therefore networks and industrial clustering. Preliminary work on the topic however indicated that much of the work on search and ways of handling enormous amounts of information was concerned more with final areas of use such as marketing and in particular search engine optimisation. Other routes of research in aiming to extract information or indeed knowledge against a set of input criteria were concerned mostly with information retrieval for libraries, the military, medical research and similar where the ability to extract something approaching knowledge from large amounts of web derived data was paramount. Cases involving industry and firms were few and generally used as exemplars for some particular facet of information retrieval technique rather than research of interest on the topic in its own right. A large part of the early research into web topics therefore became a case of finding research activity that came somewhere near what was required but without finding any related to information extraction that could make an immediate contribution to industrial cluster research. The web topics for research were thus refined into the headings of; web content mining, web structure mining and web usage mining although the latter was quickly discarded.

As a result of this a small number of appropriate papers were found where the authors had used techniques in a largely unrelated application but which could be used with some modification for research into the activity of a corpus of firms and on a large scale.

What was useful were the prior experiences of researchers in trying to deal with generic internet problems such as the presence of various kinds of ‘noise’, embedded links, html instructions and sundry references unrelated to the topic being searched.

The basic conclusion from this part of the literature survey into web topics for the enhancement of data sources or information on industrial clusters was that no substantive work existed at the time and subsequent research would have to be targeted at filling this gap in knowledge.

A similar position existed with a preliminary look at how the web might be able to identify linkages between firms. The links between firms were seen as part of the clustering process and to be able to identify them would support research which would normally be done by expert opinion or by survey, both acknowledged as time consuming processes with possible errors of omission regarding the population of firms under study. This part of the literature search was also concerned very much web structure and as with web content mining it appeared that existing research had been very much focussed on understanding the underlying structure of the web and examples of this again tended to be unrelated to anything to do with industry. At the end of this rather unsatisfactory process it was clear that to find links between firms, particularly on a regional scale it would be necessary at some point to design a program specifically to do this.

Networks and Network Theory

The review on networks started with a historical look from the earliest ideas, not because it was known that these would necessarily be appropriate but because the development of network theories have followed a piecewise trajectory with some brilliant mathematicians advancing the science followed by many years where little activity took place to be followed in turn by some new breakthrough. In some respects this mirrors in time the writings of say Adam Smith, Alfred Marshall and Michael Porter. In the same way that Marshall's ideas have come to prominence as a basis for more modern investigation on clusters so it was felt that early network theories should be considered.

The literature review in this section was therefore concerned with both underlying concepts and practical applications. A common misconception is that the network is little more than a series of joined nodes portrayed as a diagram. The review has shown that this is not so and that networks have properties which can be used to both describe and explore the behaviour of the network under defined conditions. In many of the classical definitions of an industrial cluster, in addition to the capabilities of the members the interactions between them are key and indeed this is one of the reasons why networks as a subject were looked at in the first place. As one of the desired outcomes of the literature review was to see how network theory applied to links between firms and in particular the part that could be played by the internet in discerning information about the presence and properties of such links it would appear that the literature was more helpful than in the section that merely looked at links in isolation. As always, putting the theories outlined in this part of the literature review into practice in the context of clustering may of course indicate gaps where research will need to be designed but at this stage in the investigation it was not known what those areas might be.

To summarise, it may be said that much of the work done on any form of knowledge extraction or the discernment of networks has only involved the use of the internet for applications that do not include industrial clusters directly. In the research that follows it will be necessary to see to what extent the prior research noted in this chapter can be adapted for use and what new methods will be required to be developed. The lessons learned from similar endeavours in dealing with the mass of internet derived information, the presence of large amounts of extraneous material or noise due to the fact that the internet is unregulated and non-standardised, can thus be dealt with having the benefit of what has gone before.

.

Chapter 7. Examination and Development of Tools and Techniques

7.1 Introduction and Basic Road Map

In Chapter 4 a number of possible routes to finding web derived text and associate tools were discussed. These can be categorised into three broad headings being:

- Manual methods
- Spider based text grabbers and lexical analysis
- Sophisticated programs that mimic human intervention

This section describes the results of the investigation into the three approaches above and which can be further classified into:

- Existing tools which can be used either stand alone or in conjunction with others
- As above but which require some modification
- New developments

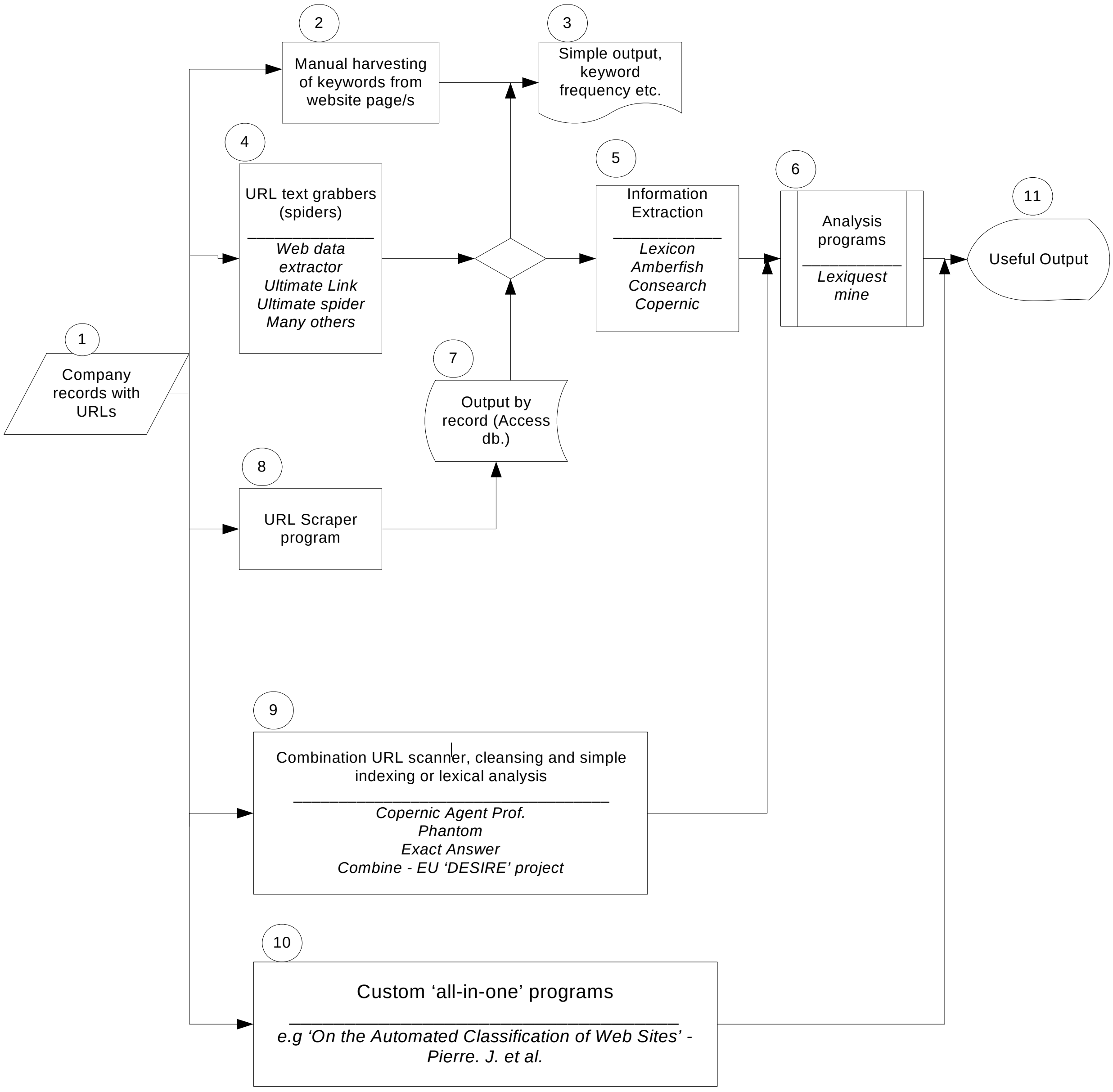
To do a comprehensive search and appraisal of all potential tools is problematical as new tools and applications are being developed all the time, often for applications far removed from the area of interest here but which could still be useful. What follows is representative of the main classes of tool. The relationship of these and how they fit into the three broad categories as above are shown in Figure 3.

For all the main elements as shown and which are discussed below, the starting point is a database of existing valid URLs. This is a reasonable starting point for research into information gleaned from individual organisation's websites although the acquisition of large numbers of appropriate URLs is a significant task that is discussed in Section 7.5, Chapter 8 and Appendix 8. Chronology of URL database build. In addition there are indications of more sophisticated methods of harvesting data and hence information that can be used which do not require such a corpus of URLs, merely some starting or 'seed' URLs. Such methods have within their programming the intelligence to find 'more of the same' by a combination of pattern matching and linking but such programs seem few in number and difficult to operate in the context of industrial cluster research. In any case this section of the research, whist important in its own right is also a necessary precursor to more sophisticated methods of information retrieval.

After discussions on the various elements shown in Figure 3 following an interim conclusion on the use of existing tools, the chapter continues with the motivation, design and results in the use of a new ‘URL Scraper Program’. The development of this program occupied a significant amount of time and effort and although the program itself was not particularly efficient from the point of view of acquiring web derived company information the exercise did serve a useful purpose in understanding some of the pitfalls in undertaking this type of task.

The chapter then continues with a summary of the research process using the most promising of the commercial programs found (Phantom) modified as required to address the research question. The chapter concludes with a comparison between the URL Scraper results and Phantom output with a further comparison between data derived from a credit reference database and textual output from the Phantom program.

Figure 3. Process Flow Diagram


1
Company records with URLs
2
Manual harvesting of keywords from website page/s
3
Simple output, keyword frequency etc.
4
URL text grabbers (spiders)
Web data extractor
Ultimate Link
Ultimate spider
Many others
5
Information Extraction
Lexicon
Amberfish
Consearch
Copernic
6
Analysis programs
Lexiquest mine
7
Output by record (Access db.)
8
URL Scraper program
9
Combination URL scanner, cleansing and simple indexing or lexical analysis
Copernic Agent Prof.
Phantom
Exact Answer
Combine - EU 'DESIRE' project
10
Custom 'all-in-one' programs
e.g 'On the Automated Classification of Web Sites' - Pierre. J. et al.
11
Useful Output

Explanation of the main elements in Figure 3

Starting point - #1

The process above starts with the Author's own database of 4300 N.E. regional companies each with a published URL. The provenance of these URLs is known in that they were manually checked in 2002 by a small call centre with the aim of selling them on for marketing purposes.

In some respects the URLs used do not need to be ordered or representative of any particular sector. They are being used initially as a 'test bed' to feed into the various analysis programs. Clearly it is better if they are up to date and there are few broken links. Further work has been done on separate analyses to establish comparators including an SIC profile of companies in this and other samples. This is reported on in Section 9.4.

Manual Harvesting - #2

This is the simplest process but also prohibitively time consuming and hence expensive for any database over say a few hundred records. The basis is that a human operator looks up the URL, moves around the web pages making notes on what are deemed the key features, markets and operations of the company being examined. At this point the operator could compare his or her judgement as to what the company does, write a summary to the original database and even 'classify' the relevant words or phrases back into SIC categories.

Even for a small region such as the North East of England with 42000 VAT registered businesses in 2004 and with approximately 40% (in 2003[20]) having a web

[20] The full database used here contained over 13000 records of which approximately 40 % had what appeared to be a valid URL. However some 900 were found to be non-company related and referred to for example individuals. This is a recurrent problem with other URL databases and provenance of the data here and other subsequent data sources are discussed in Appendix 8. Chronology of URL database build.

The assumption is that the 40% figure can be applied to the general company population but this may not be strictly accurate. Biases of this nature are discussed also in the Appendix and elsewhere in the text.

address, a coarse estimate of the time to undertake such a task would be several person-years.

This process is useful however for comparing a small number of records processed manually with the same set subsequently processed without human intervention. There is of course the issue of individual bias, particularly if the task was split between a number of people. Even with guidance notes it would be difficult to standardise the methods of gaining keywords and descriptors on the basis that some investigators might give greater credence to some aspects of a particular website than others looking at the same site.

Text harvesting tools and Spiders - #4

There are number of tools in this category available through academic, shareware and commercial sources. A good example of the latter is 'Web Extractor' [www.webextractor.com] which does more or less as advertised in that given a database of URLs it will interrogate each site to a specified depth and return the metadata, email, telephone and fax numbers. The 2004 version however did not return body text. This is a major shortcoming in that, as will be reported later, only between 14% and 30% of sites, dependent on sector, have metadata within them. The result (text) however is written into a format suitable for export into a database for further analysis. Given a set of metadata this can be used in the next category of processing – information extraction.

More sophisticated programs on the same basic theme are Lencom's Robonavigation Office 4 which is a suite of programs including Visual WebTask 4. This latter has an impressive simple search capability [visual-web-task.lencom-software-inc.qarchive.org/]

Most of these programs are set up to gain web related information on a single topic and then to use a domain specific and tailored search engines. For example if the unique word 'Cullercoats' is submitted to WebTask 4 it returns a large collection of references from many search engines and in a format suitable for tracking the source.

In addition to the possible forms of this type of tool discussed above Van Gemert (2000) looked at 64 text mining tools and published a brief summary of the capabilities of each. Where available the source and costs were also given. These programs were also examined in the context of the research question i.e. could they be used to contribute to our knowledge of industrial clusters. For the most part they fell into two categories being either clearly inappropriate or inaccessible (usually for reasons of acquisition cost even at academic rates when this option was available).

Information extraction #5

Information extraction is a significant topic and refers to programs that automatically extract summaries of large tracts of text such as that to be found in libraries, patent applications or theses._An example is Copernic Summariser [www.copernic.com] which can take coherent text or a series of phrases, for example words extracted from body text and then summarise what it thinks are the salient points. It does not cope well with a collection of metatags and the commercial product can only do one text at a time (although that text could be very large). Other products in the same category are;

- Amberfish {www.searchtools.com/tools/amberfish.html} for text searching and indexing although this is probably better in box #3
- Lexicon {http://www.funsci.com/texts/index_en.htm#lexicon} which is a simple lexical analysis program. Both these products are freeware.

All of the above have some difficulty operating ‘off the shelf’ or in combination with other parts of the process. As has been mentioned before, they were written primarily for an application other than the one of interest here.

Although the search for suitable products in this category continued, at this point in the investigation a decision was taken to design and write a program that could do what was required, at least in terms of relevant text extraction. In a reflection of the way this program functions it was termed a ‘URL Scraper’ and is described in Section 7.2.

Analysis Programs #6

Analysis programs are the next level of sophistication from relatively simple information extraction. Dependant on capability, analysis programs can take either raw text or output from an information extraction engine and try and derive relationships and meaning understood by humans. These programs for the most part (but not exclusively) fall into the general category of lexical analysis and employ data or text mining technologies.

They range from the relatively simple (and free) such as Lexicon and Wordsmith tools up to more sophisticated and expensive proprietary tools such as Lexiquest Mine.

Although clearly it is possible to try the worth of freeware, extensive ‘try before you buy’ tests on expensive commercial software are not practical and assertions from the vendors that their product would ‘be ideal’ may not in practice be true.

Combination programs #9

These programs aim to be a seamless combination of the other independent elements discussed earlier, i.e. URL scanner, cleansing and simple lexical analysis.

These programs are again variable in their capability with a number of low cost proprietary offerings at one end to an EU collaborative effort known as the DESIRE[21] project at the other.

The proprietary programs examined have been written to solve a wide range of problems in extracting information from the web but still seem to suffer from at least some of the problematical features of independent programs. In particular the ability to remove ‘noise’ is a key feature and without such capability any subsequent processing, no matter how sophisticated, will struggle to discern useful information.

[Ghani *et al* [(1999)] summarised the problem as follows:

“Information extractors and classifiers operating on unrestricted, unstructured texts are an errorful source of large amounts of potentially useful information, especially when combined with a crawler which automatically augments the knowledge base from the World Wide Web.”

There is a further class of combination programs, which function in a different way to those above and which hold considerable promise at least at the keyword level. PHANTOM from Aktiv Software Corporation is one such. Phantom works by

[21] http://www.desire.org/ he DESIRE project ran from July 1998 until June 2000 and was a collaboration between project partners working at ten institutions from four European countries - the Netherlands, Norway, Sweden and the UK. The project's focus was on enhancing existing European information networks for research users across Europe through research and development in three main areas of activity: Caching, Resource Discovery and Directory Services.

setting up a spider according to rules set by the user. These rules control depth of search, the links to follow – internal and external and type of text to acquire. The text search can access title, metadata, keyword, html tags and body text. It can also handle Word and pdf files and convert them to text on the fly.

For a given database of URLs the spider will identify web pages containing useful text and write them to a datastore on a local disk. The program then indexes the pages excluding a list of program defined 'noise' words such as 'the' , 'of', 'is' etc. The user can add and modify the noise words to be excluded.

In order to gain results the user inputs a keyword or phrase of interest or implements a Boolean search, the relevant web addresses together with all the text found and stored on the site is shown. As all the pages are stored locally (on the user's machine) the process for doing this is very fast.

Test results with this program have been good and are discussed in the next section.

The program however was written for a use on a single large site or a small number of related sites with the same top level domain such as a University or large industrial group.

The downside therefore is that for a multilevel search on a corpus of URLs the time taken and the size of the stored file can be large. The time taken is a function of the number of records searched, the speed of internet connection and the response time of the site being spidered. There is also an internal programming restriction that limits the number of URLs that can be searched in any one session to < 1000.

Custom Programs #10

This brings us to the final category in which it is acknowledged by researchers in the field that currently there is no single program or combination of proprietary programs that can take either a corpus of URLs or some other starting point for a spider and then return useful knowledge (as distinct from data).

As a result there are number of investigative teams with expertise in machine learning, artificial intelligence and computer science who have worked on custom programs.

Two examples of this are the one by Ghani et al (1999) already noted and Pierre (2001).

The former was put together on the basis of 'wrapper induction'. The wrapper is a descriptive form of words in the record whose URL is subsequently scanned. An example already given is of the Hoovers Online Web being a detailed resource of a large number of companies. The wrapper is used to extract features which are then used to ‘train’ extraction algorithms applied to ‘noiseful’ web derived text or data.

In the paper by Pierre the author started with a set of company URLs with a known NAICS classification. In this highly relevant paper entitled ‘On the Automated Classification of Web Sites’, the aim was to write a program that could access text relating to the companies from their websites, filter out noise and build a knowledge base that could then, with some statistical processing, identify words and phrases that could be then correlated with existing NAICS descriptors of individual firms. The NAICS numbers thus deduced were then compared with the original NAICS as given by the companies.

In many respects the strategy adopted follows the form of the aims here but with some important differences. Principally we are trying to seek out additional activities that the firm undertakes and it could be argued that matching given NAICS numbers with ‘derived’ NAICS numbers is contrary to that aim. Ideally the program should find and match more than one NAICS or other SIC categories if appropriate for any particular URL. It also makes the tacit assumption that the NAICS numbers given in the original data are ‘correct’, at least for the purposes of comparison with NAICS numbers using text derived from the websites. As noted previously a number of commentators have criticised the process whereby the accuracy of many SICs may be found wanting in the context of an individual firms activities. This seems to be due to persons within the firm being asked carry out a precise categorisation of their firm’s activities but for a variety of reasons do not do so with any degree of consistency. It is largely a problem of quality control but apart from the data collecting authorities urging firms to take some time and care over their choice there is probably not much that can be done about the situation. Thus in the case under consideration here the degree of comparison between given and derived NAICS numbers found by Pierre could be better or worse than the results found because we are dealing with two variables rather than assuming that the given firm NAICS number is a constant.

7.2 URL Scraper Program #8

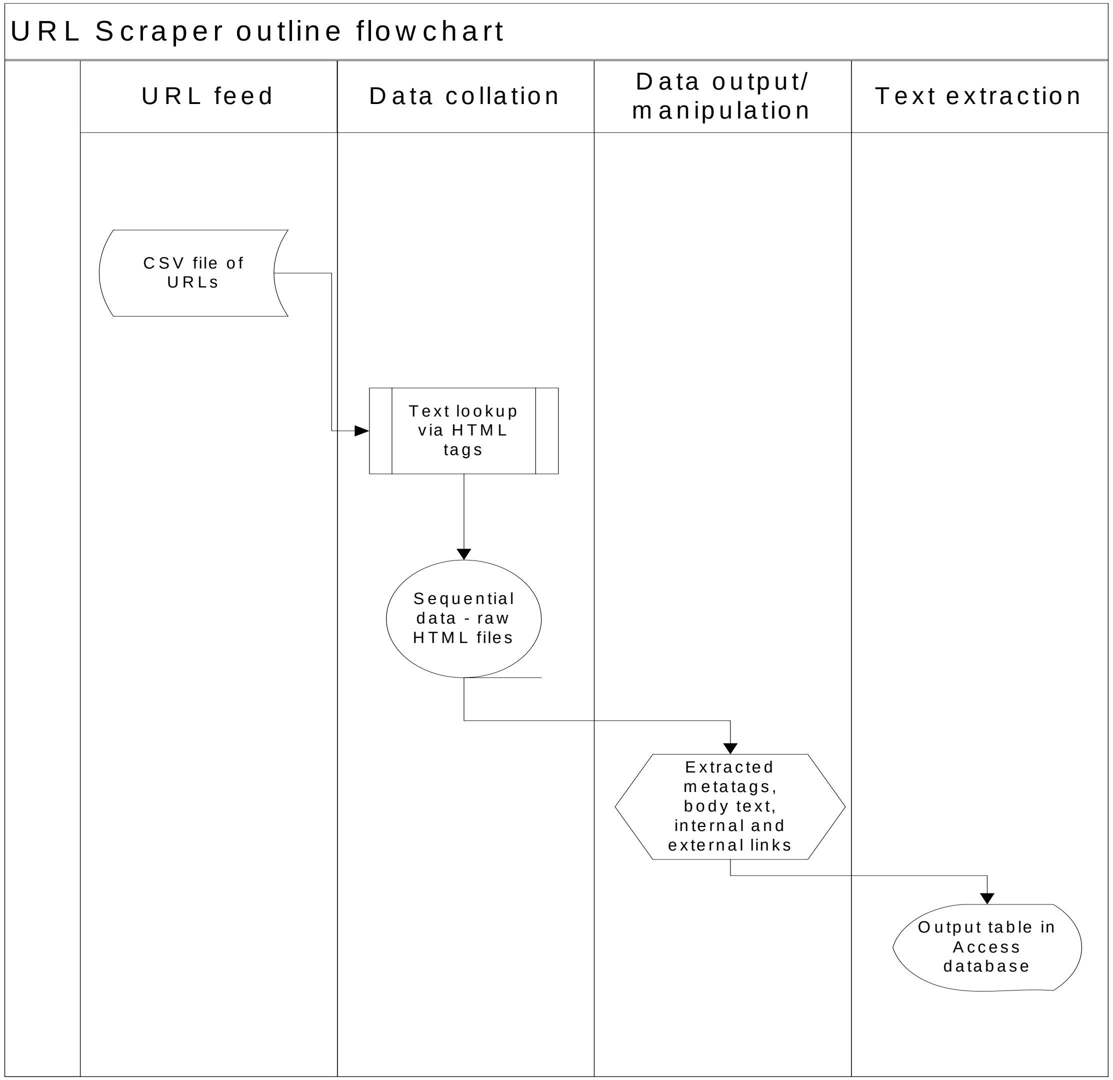
The motivation for writing a routine to employ a spider to access useful URL derived text was twofold. Firstly, at the time of writing it appeared that few if any commercial, freeware or other available program existed that appeared to provide all the features that might be required in order to carry out meaningful research. Those few methods and/or programs discussed that might offer something useful (with the notable exception of the PHANTOM program), are proprietary and are either not available or are too costly to test, there being no guarantee that they would be suitable in any case. The URL scraper program clearly has no such restrictions.

Secondly in the context of gaining first hand knowledge of the possibilities and limitations of spidering web derived text it was felt that it would be a useful exercise even if the experience turned out to be negative to some degree.

The specification for this program and outline flow diagram is shown in Figure 4 below.

In simple terms the program takes a business database in csv format with a URL field, identifies and validates the URLs, interrogates individual company websites against criteria set initially by the user and for each page on the site strips out the metadata and body text. It also reports links found both within the site and also references to external sites. The program then writes the results into a standard Access database for further analysis.

Figure 4 URL Scraper outline flowchart


URL Scraper outline flowchart
URL feed
Data collation
Data output/ manipulation
Text extraction
CSV file of URLs
Text lookup via HTML tags
Sequential data - raw HTML files
Extracted metatags, body text, internal and external links
Output table in Access database

A series of tests was designed to prove and tune the program and to gain knowledge on how programs of this nature function in practice.
These are discussed in the next section.

7.3 Findings

These findings relate to the testing, where possible, of the some of the routes outlined in Figure 3

7.3.1 URL Scraper results

The program can take as input single URLs and also records in a database. The single URL feature is useful in that it allows the user to test the efficiency with which the program grabs individual web pages and strips out text, either from metatags or from body text or from both sources. Efficiency here is a term that can be defined as the ratio of program derived data (metatags and text) 'scraped' divided by the actual text on the page. The term 'actual data' refers to both visible and hidden text such as metatags. It became clear that the output from any group of SIC determined URLs is very uneven with some sites returning little in the way of useful text with others loading up the output database with masses of text from multiple pages. In addition, apart from the timeouts on URLs not found, the program would lock up on certain types of URL and could take anything up to 6 minutes to completely search and 'scrape' a single large site. Examination of the resulting Access database showed multiple entries and a lot of extraneous noisy text and data. Much of this noise is in the form of HTML tagged material put in as part of the control of the site or genuine messages but which were not useful in the context of company descriptive material. Examples of the latter are downloads for various plug-ins such as RealPlayer and Adobe Acrobat readers. This was known about when writing the specification and flow diagram for the program and as such factors needed to be addressed the program had been written with filters to take out or to ignore such returned material. It was however not so much the scale of the irrelevant data but the variety that was unexpected.

During the test phase, in order to keep the run time down small numbers of records were fed into the scraper program for each test i.e. < 200. It was found that the percentage of sites having broken links, metatags, body text or both was a function of the quality of the database. In addition it was also found to be highly sector dependant.

As an example for the results of searches, Table 1. SIC metrics by group gives an indication of the metrics associated with returned text for different SIC groups.

Table 1. SIC metrics by group

UKSIC	Description	% scraped	% metadata	% body text	% both	Comments
51	Importer/exporters	63	25	13	13	
51.13	Builders Merchants	80	30	10	20	
55.3	Restaurants/cafes	45	9	18	0	very low strike rate
60.24	Haulage Contractors	76	21	43	9	
72.2	Computer systems & S/W	79	9	27	54	
72.6	Computer Services	71	12	35	35	
74.14	Business Services	69	7	13	9	
74.84	Engineers - various	87	41	54	26	many multiple outputs
74.11	Legal services	93	6	13	10	
74.12/3	Accountants/marketing	91	10	36	54	
	Averages	**75.4**	**17.0**	**26.2**	**23.0**	

The following is noted:

In the database list of URLs, on average 75% were opened by the program. Of these only 17% contained metadata with 26.2% gaining some sort of body text with 23% of sites showing both body text and metadata.

However the individual groups of SIC based sites showed a wide variation in individual figures with Business Services being as low as 7% on metadata and 'Engineers' as high as 41%. These figures refer to some sort of text entry in the two categories as above. This does not necessarily mean that the text provided useful knowledge.

Nevertheless it became possible, in spite of the very obvious shortcomings of both the program and the data sources to discern some interesting results and a way forward.

The output from the Scraper program writes into an Access database and is therefore searchable by for example a nominated keyword. For the SIC group 74.84 'Engineers- various' there are no references in the SIC description to activity associated with say 'offshore' engineering. A word search however reveals text references to 'offshore' in 7 companies out of 179 in the cohort. Similar results are obtained from 'defence', again an SIC that hardly appears in the region's companies but it is a matter of record that there are over 250 members in the regional trade body the Northern Defence Industries. This organisation exists to promote trade in defence related markets so presumably its member firms are similarly interested in such markets.

In this respect at least the program indicated a potential for useful knowledge that might be used as part of a cluster study.

7.3.2. Conclusion and Further Work on URL Scraper

The presence of large amounts of noise words and duplicates, whilst irritating, is not necessarily a major issue for word searches or short phrases on specific topics although care must be taken with any metrics based on keyword density because of multiple entries on the same site. However if one is seeking to extract knowledge or answer specific questions the highly unstructured nature of web derived text does become a problem. As mentioned in Custom Programs #10 other researchers such as Pierre (2001) have attempted to cope with this by a combination of machine learning and statistical processing.

The main lessons learned from what was a significant effort to get the Scraper programme working with even a modest degree of efficiency are:

1. The results derived are at best modest and a high degree of effort in program testing, modification and subsequent redesign was required to get to the stage where it was felt that something useful was being learned.
2. The presence of residual noise in the outputs was a continuing problem and whilst simple keyword searches were not affected by the presence of such noise it detracted from the possibility of subsequent processing to either parse text or to gain knowledge by some other means on a hands free basis.
3. Any program can only pick up what is available. It was noted in the case of metatags that whilst their use can be a useful pointer for search engines, by manual inspection only a minority of sites actually made use of metatags. This may be because not all search engines take account of metadata and also in the recent past Google has 'punished' website authors making excessive use of metatags to gain some unwarranted advantage in the ranking of their site. Therefore whilst in the matter of keyword search, metadata is useful here on the basis that authors put into these tags words that summarise the activities of the organisation, in practice the actual presence of metatags on sites at about 17% indicates that it is a minority source of information.

As an independent check a metatag analyzer [http://www.submitexpress.com/analyzer/] was used on a sample of sites to check the presence of valid metatags. The results, not surprisingly were similar to those of the URL scraper program for the same corpus of firms but as has already been noted the proportion of sites with metatags is very sector dependent.

7.4 'Phantom'

This program [http://www.phantomsearch.com] was tested with the same data sets as used for the URL Scraper program. It was thus possible to have a direct comparison between the URL Scraper which basically grabs all text that it can find on given web sites and then places this text in an Access database for either keyword search or further information extraction.

Phantom, as described briefly in the section on Combination programs on page 103, operates on the basis of taking the base URL and then searches for available text and other valid 'words' on pages to a user specified depth. The text information on the pages is extracted and written to a local hard drive where it is then indexed. Subsequent searching of the index is therefore very fast. As web pages containing the title, metadata, keywords, html tags and body text, text conversions of pdf files and MS Word documents are indexed a rich vocabulary of company descriptions and activity is obtained. A key feature of Phantom is that within the index there is a hyperlink to the original site from which any particular word of interest was originally extracted and a relevant summary of the company's activities is shown in the output. An example of these is shown in Figure 5. In the first full test the database of 'Engineers- various', with 179 records was used with folder depth 1 with links followed. Some 149 URLs were examined by the robot and 1884 pages were indexed. The process took 129 minutes with the index search revealing 26 companies with the word 'offshore' somewhere on their website. The comparison with the Scraper program for the keyword used as an example is shown in Table 2 below.

Table 2. % of sites indexed by ‘Phantom’ & URLScraper

Keyword	Offshore'	
COMPANY	**Phantom**	**URLScraper**
ATA	y	y
Handmark	y	y
Northview	y	y
Wyko	y	
Elfab	y	y
Inlec	y	
Mechtool	y	
Furmanite	y	
Econnect	y	y
Lawson	y	
K-Home	y	y
Eant	y	
Dowding & Mills	y	
Cylinder	y	y
Ovalway	y	
Entecuk	y	
Bel	y	
Amec	y	y
Contract Design	y	y
Hogg	y	y
Allison Hydraulics	y	
HSE	y	y
Inlec	y	y
MKW		y
Percentage of total sites visited (170)	13.40%	7.80%
% of URL sites indexed/scraped	72.60%	87%

Of the percentage of the number of sites yielding text as a proportion of the total URL sites in the original database it would appear that the Scraper program with 87% is the more efficient. However closer examination shows many errors, HTML tags and useless text messages in the output database. This did not happen to any great extent in the Phantom Program.

The results for selected keywords however showed the Phantom Program outperforming the URL Scraper almost 2 to 1. The keyword 'Offshore' was chosen as it is known that even as late as 2003 the North East Region enjoyed a thriving construction industry related to Offshore platforms and support equipment to such an extent that it was regarded by the regional development agency and others [Williams (1998)] as fitting with one of Porter's definitions of an industrial cluster. However examination of the SICs (SIC 74.84) of the companies used for the URL search gives no indication at all that many of these companies were involved in the offshore sector. Indeed, 'Offshore Oil & Natural Gas Services' (SIC 11.1) has relatively few entries and conventional cluster studies of the region [Miller *et al* (2001)] have relied upon text descriptions from Dun and Bradstreet business data to try and gain some knowledge regarding these 'hidden' areas of activity such as the offshore sector.

A second test involving 'Business Services' (SIC 74.1*) was carried out. As well as solicitors and a variety of consultants and marketing firms, the business services SIC has often been used a dumping ground for all manner of services that do not appear to fit easily into any other SIC category. As such it was a good candidate group for seeing what could emerge.

Both programs had difficulties with the cohort of 252 starting URLs. The URL Scraper crashed frequently and Phantom seemed to spend an inordinate amount of time accessing pages on some obviously large sites. In particular some of the large (and global) consultancies such as *www.PWC.com* (Price Waterhouse Coopers) and *www.Deloitte.com* are massive sites with upwards of 20000 pages. As another example RS Components had at the time a published catalogue in excess of 300,000 products across 89 countries and www.vishay.com a similar electronics components site had amassed some 16085 pages when it was stopped after 13 hours. Clearly attempts to fully index sites such as these as part of a wider search is asking for trouble but at the time when use of the programs were very much at the exploratory stage it was not obvious which sites were large and that unrestricted following of links would cause such problems.

There were a number of other sites also that looked like dominating both the pages scanned and hence the subsequent index. The decision was taken to remove the three very large sites noted above and a few others and to concentrate on sites with URLs relating to genuinely regional firms. Even so with an input file of 241 URLs, Phantom took over 8 hours to examine 199 sites, an 82.5% strike rate. Because of the link following feature some 12260 URLs were processed and 5728 were indexed from 12325 pages. File size was 93.3 Mb.

Such depth gives a very rich corpus of material on which to carry out searches, either by keyword or phrase, the downside being the time required although once the search and local indexing was complete all subsequent searches were very fast being accomplished in a few seconds. To give an example of the power of this, law firms often specialise but this is rarely evident to any great degree from either the SIC or the business description. A search on Phantom for ‘human rights’ OR ‘immigration’ showed 6 regional solicitors and 4 other organisations involved with this speciality.

In Phantom the program can control the depth of folder search. In these preliminary results as above the program was run on maximum folder depth with all links followed and also other parameters were set to index URLs outside pages referenced from a valid page. It was therefore perhaps not surprising that the whole process took so long. As more experience was gained tuning the program’s capabilities it became possible to achieve a balance between depth of search and reasonable time of search process.

In some respects the above runs performed are preliminary and still in the test phase but as more experience with the use of these tools was subsequently gained so was a better understanding of how useful knowledge might be gleaned from the web, balanced with an extended running time for the extraction programs.

7.5 Full search of 4300 URLs using Phantom

7.5.1 Preliminary Remarks.

At the time of undertaking this part of the research (2003-4) the author had access to the 4300 URLs already noted. It was felt such a number was large enough to test the concepts but clearly was still only a sample of the regional stock of businesses. The sample also, whilst being valid in terms of the proportion of live links, was initially an unknown regarding the representation of SICs compared to the regional stock of businesses. More precisely it was not known how a particular company had been selected by the call centre in order to verify that company's details and thus it was not known if the 4300 companies used were in any way a fair representation of the regional stock of businesses. Unless the proportion in the sample with a URL reflected the SIC profile of all regional companies then there would likely be some bias. For example certain types of firms would be more likely than others to have a website e.g. ICT firms.

Nevertheless in order to prove the concept of at least keyword and possibly knowledge extraction it was felt on balance that the 4300 URLs were indeed a good starting point although it is fully acknowledged that it would be a case of the more URLs the better in subsequent investigations. The above remarks however only apply to finding interesting words and phrases. For a robust comparison of the worth of the proposed methods an understanding of the SIC profile would also be a requirement. This would be necessary in order to undertake any comparisons on the basis of any industrial 'activity spectrum' derived firstly by SIC based data and secondly from that derived from programs of the type under discussion here.

An attempt at SIC profiling was undertaken albeit at a coarse (2 digit level) and this is shown in Appendix 5 SIC Profiles for the 4300 regional organisations.

Initial Limitations

The difficulties noted previously with a full depth unrestricted search and indexing of a few hundred URLs precluded the search and indexing of the entire URL database. However by just indexing the top level (no links) pages it was possible to undertake this task. A number of duplicates were removed together with erroneous addresses giving 3956 as the number of starting URLs. Because of internal program limitations (32k input size maximum for a .csv file) the task had to be split into four sections with the resulting indexes subsequently searched as one. The result of this exercise was encouraging giving a reasonable understanding of the activities of a larger cohort of regional firms.

7.6 Output from customised Phantom program – a comparison

7.6.1 Basis

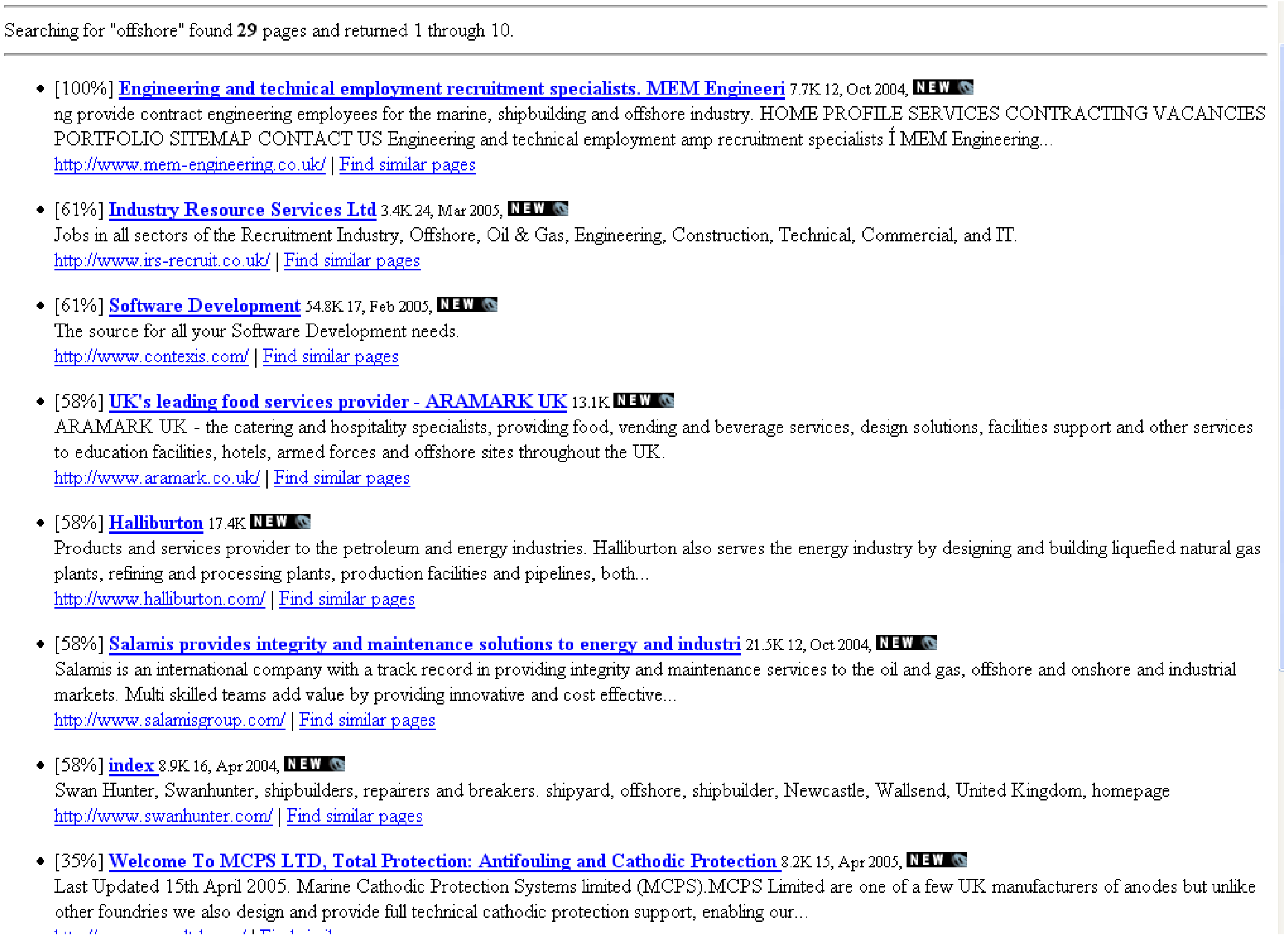
The form of output from the Phantom program related to the search terms is as shown below in Figure 5.

As can be seen a short text summary related to each of the indexed URLs is given and the links are 'live' in that they are clickable back to the original website. This facility gives an instant check that the URLs is both valid and relevant to the search term or terms entered[22]. For the search term 'Offshore' entered the program immediately finds any URL in the database with that word anywhere on is website.

It is clear that whilst the output from Phantom looks promising in isolation some direct comparison with more conventional data sources would give a useful indication of the robustness of the information gained. An obvious way to do this would be to compare this with readouts from a conventional firm database such as one from DNB. It is important however to design the comparison with the objectives of the research in mind. We are trying to find ways of acquiring or augmenting information about the firm and its activities, not necessarily attempting to find a complete substitute for conventional sources.

[22] Technical Note – The screen shown is an example of the output from Phantom. However the database being used here has reduced number of OCW derived entries being just 1703. The reasons for this are due to continuous development of the URL database and the subsequent removal of duplicate entries and which is explained in Appendix 8. Chronology of URL database build.Consequently only 29 entries are shown for a search on 'Offshore'. The form of output in Figure 5 is however independent of the number of URLs indexed and an any case the figure of 29 entries rises markedly with additional URL data obtained through the course of the subsequent research.

Figure 5 - Output screen from Phantom Search Program



Searching for "offshore" found **29** pages and returned 1 through 10.

- [100%] **Engineering and technical employment recruitment specialists. MEM Engineeri** 7.7K 12, Oct 2004, NEW
ng provide contract engineering employees for the marine, shipbuilding and offshore industry. HOME PROFILE SERVICES CONTRACTING VACANCIES PORTFOLIO SITEMAP CONTACT US Engineering and technical employment amp recruitment specialists Í MEM Engineering...
http://www.mem-engineering.co.uk/ | Find similar pages
- [61%] **Industry Resource Services Ltd** 3.4K 24, Mar 2005, NEW
Jobs in all sectors of the Recruitment Industry, Offshore, Oil & Gas, Engineering, Construction, Technical, Commercial, and IT.
http://www.irs-recruit.co.uk/ | Find similar pages
- [61%] **Software Development** 54.8K 17, Feb 2005, NEW
The source for all your Software Development needs.
http://www.contexis.com/ | Find similar pages
- [58%] **UK's leading food services provider - ARAMARK UK** 13.1K NEW
ARAMARK UK - the catering and hospitality specialists, providing food, vending and beverage services, design solutions, facilities support and other services to education facilities, hotels, armed forces and offshore sites throughout the UK.
http://www.aramark.co.uk/ | Find similar pages
- [58%] **Halliburton** 17.4K NEW
Products and services provider to the petroleum and energy industries. Halliburton also serves the energy industry by designing and building liquefied natural gas plants, refining and processing plants, production facilities and pipelines, both...
http://www.halliburton.com/ | Find similar pages
- [58%] **Salamis provides integrity and maintenance solutions to energy and industri** 21.5K 12, Oct 2004, NEW
Salamis is an international company with a track record in providing integrity and maintenance services to the oil and gas, offshore and onshore and industrial markets. Multi skilled teams add value by providing innovative and cost effective...
http://www.salamisgroup.com/ | Find similar pages
- [58%] **index** 8.9K 16, Apr 2004, NEW
Swan Hunter, Swanhunter, shipbuilders, repairers and breakers. shipyard, offshore, shipbuilder, Newcastle, Wallsend, United Kingdom, homepage
http://www.swanhunter.com/ | Find similar pages
- [35%] **Welcome To MCPS LTD, Total Protection: Antifouling and Cathodic Protection** 8.2K 15, Apr 2005, NEW
Last Updated 15th April 2005. Marine Cathodic Protection Systems limited (MCPS).MCPS Limited are one of a few UK manufacturers of anodes but unlike other foundries we also design and provide full technical cathodic protection support, enabling our...

The comparison was therefore designed on the basis of the following two precepts:

1. Look at comparative counts i.e. are the numbers emerging from the two different methods at all comparable.
2. Look at direct comparisons with existing data at the individual firm level

With respect to the source of the data used for the comparison the author had access to TrendSCAN and the Trends Business Research Ltd database[23]

7.6.2 Method

TrendSCAN is analytical software containing information on approximately 2 million UK businesses and organisations. The version used here was populated by data from 1999 and 2002, enabling analysis of change between these dates. Although a longitudinal study is not part of this investigation a key feature of TrendSCAN is its ability to undertake such studies and is one of its main selling points. The economy can be analysed by region, county, district, ward and postcode. Markets can also be analysed by descriptors of business activity supplied by the firm although these descriptors vary in length from a few words to a few lines of words. The product is also configured to provide 5-digit SIC, sector and supply chain analysis.

For the purposes of this study however the TrendSCAN software has great merit in that it allows fast counts against user specified criteria and also allows interrogation of selected fields of data for each company record.

The first objective, to look at comparative numbers was not a simple matter of comparing say sets of SICs derived from TrendSCAN with those from Phantom as Phantom specifically does not elicit SICs directly, indeed the fact that it can derive a rich text of firm activity without recourse to SICs or SIC descriptors is one of its strengths. Additionally there is a wide disparity in the number of firm populations for each dataset.

The initial Phantom dataset for the North East Region was 4300 URLs (companies) whereas TrendSCAN is based on a (regional) population of some 20000.

[23] TrendSCAN is a trademark of Trends Business Research Ltd (www.tbr.co.uk), a business consultancy that has access to all 2M+ annual UK company records from DNB covering many years.

As an example at a more detailed level, in ‘UKSIC(92) 74140 – Business and Management Consultancy’ TrendSCAN returned some 494 regional entries whereas Phantom gave 54 firms when the term ‘management consultants’ was searched.

Clearly unless some means of closely matching the respective company populations, both in numbers and broad SIC profile, could be found there was little point in continuing with comparative counts based on SICs or indeed anything else.

Attention was then given to the second objective in this test which was to look at direct comparisons with existing data records at the individual firm level. This was done by choosing companies that appeared on both the TrendSCAN and Phantom databases. Initially this was looked at in a somewhat arbitrary fashion to get a feel for the information coming out of each database as regards company descriptors. It should be reiterated that TrendSCAN data contained, in addition to the UKSIC for each record, a text description of company activities although, as noted, this varied from a few words to several lines.

It has been noted previously that certain SICs are difficult to deal with as they can be so general as to admit a wide variety of firm activity and occupations. Examples are those associated with business services and anything with the suffix ‘n.e.c.’ being ‘not elsewhere classified’. Two examples of these codes are:

74000 – Other business activities

74849 – Other business activities n.e.c.

As a result these categories become populated by a wide variety of activities that are not so much unique but more the case of having being used by the originator who could not find a good match to their own company category when the original data was being collected. This topic was also discussed previously in sections 3.3 and 3.4.

Therefore much of the examination of matched firms from the two databases looked at these categories on the basis that they would provide a more robust test for Phantom. Four samples from this comparison are shown in Table 3 below with the output from the TrendSCAN data on the left column and the Phantom derived output in the right hand column. The text shown is verbatim from each respective output and has not been edited.

Table 3. Comparison of TrendsSCAN and Phantom data

TrendSCAN	**Phantom**
Entek UK Ltd. UKSIC 74140 Business and Management Consultancy Market:Professional Business Services Activity: Environmental Consultants	Entek UK Ltd. One of the UK’s leading environmental and engineering consultancies with over 50 years experience in the public and private sectors. Skills planning and environmental appraisal, water, waste, contaminated land, health and safety
Rhodia Chirex UKSIC 24140 Manufacturer of other organic chemicals, fine chemicals and intermediates	Rhodia Chirex Chemical outsourcing partners for the chemicals industry. Scale up manufacturing process development, contract research, process support
NJM UKSIC 74140 Business and Management Consultancy, professional business services, management consultants	NJM Specialists in economic regeneration and development policy. Identification of EU/UK funding opportunities, grant funding applications, management of grant funding projects
Nigel Wright Consultancy UKSIC Recruitment, Placement agency, Commercial products and services	Nigel Wright Consultancy Leading firm of recruitment and management consultants. Experience of appointing finance, legal, general management, IT, sales and marketing, operations, supply chain and HR people

7.6.3 Conclusions on test

The following were concluded as a result of this short comparison;

1. The text descriptor shown by Phantom is much more comprehensive than anything similar available from SIC based text descriptors or additional words from TrendSCAN and hence from DNB. It should also be understood that the Phantom derived text shown is a summary, the length of which is merely a feature of the program set up. The actual text 'available' to Phantom is that extracted from the whole of the company website to the depth specified in the initial search including all MS Word documents and pdf files as well as various forms of html related text including metatags.

2. Some of the activity descriptions, such as that for NJM, reveal sets of activity completely unknown to the SIC based version. This is a key point as further examination on SIC 74140 of individual firms that could be matched showed a variety of firm activity on Phantom not shown in the DNB derived descriptive text.

A further strong conclusion that emerged was that if attempting to do more comprehensive comparisons of this type it would be necessary to acquire a much greater size of database and hence a greater population of starting URLs.

7.7 Observed Benefits and Disadvantages

For both URL Scraper and Phantom it is quite possible with a simple keyword search to find both the location and the incidence of any industrial activity if it exists within the website of the database of company URLs scanned. The programs however do rely on the user deciding which keywords or phrases to search for.

This is perhaps not unreasonable because if the programs were modified to give keyword popularity or frequency of occurrence counts this would not necessarily give an accurate understanding of activity within a region. For example it is likely that the word 'internet' would appear with a high frequency, not necessarily because technology based firms were engaged in the development of internet related intellectual property but because a wide variety of firms use the internet in their business and say so on their website.

All the tools examined have some disadvantage when set to the task of automatically extracting information from the web. Some of these shortcomings include limitations on the number of records, limitations on web searching capability and most of all the presence of 'noise' when operating on what is a completely unregulated and non standardised source of mostly text based data. For those programs that do yield possible useful data that can be processed by other means the size of the pages or text fragments extracted can be very large and even with a fast internet connection the spiders used can only operate as fast as any individual site can transfer data. The net result of examining more than a few hundred sites when following links is that it becomes at least an overnight batch job. Conversely however if a program such as Phantom is used without following links then the process for even a large number of URLs can be carried out in a reasonable time. The downside is that a smaller number of pages for each site are indexed and thus some of the rich detail of deep crawls is not found. This latter aspect could be important dependant on the investigation being carried out. It should be theoretically possible to use a 'no links' search to identify areas of activity that might constitute a cluster and thus identify a smaller number of appropriate companies. With such a smaller corpus of firms and other organisations a much more detailed search 'following all links' could then be carried out to investigate with a high degree of resolution the activities of such a reduced number of organisations. One other issue that did crop up was the problem of Macromedia 'Flash' sites or sites with a Flash front end. These are essentially graphical sites in which, unless there are buried metatags, no text exists for the spider to access. The fashion for Flash waxes and wanes as although Flash sites can have strong visual impact and in particular can be made dynamic such sites give all search engines problems. Although the more sophisticated search engines can derive text and hence search metrics using pattern recognition software most web developers seeking to attract the attention of search engines look to put text easily found by the search engines and usually avoid 100% Flash. In our case the spider does not 'know' that there may be words represented as a picture and consequently, in the absence of any form of hidden text, will return 'no text' against that particular URL. Other sites that the spider found difficulty accessing were a few that required some action on the part of the user such as 'Click here to enter' buttons or those requiring a password.

Finally of course for some sites the webmaster may actively discourage unknown spiders from accessing the site, particularly during working hours as it is possible for a comprehensive spider crawl to slow down a site's response to other users. Phantom has the facility to both request access in the first place and also to adjust the time between activating requests for information. Most webmasters object to a spider 'hammering' away at their site on a continuous basis thus blocking off or at least slowing down legitimate search enquiries. Thus the time between requests by

Phantom can be adjusted to between 1 and 30 seconds. In practice as most of the sites were regional to the North East and therefore subject to local time (GMT) and as the majority of spidering was done overnight this time was set towards the lower end of this scale. If any spidering had to be done during the normal working day, for test purposes for example, then the time interval was set for longer periods.

In later chapters the lessons learned from this exercise have provided useful knowledge to carry out a wider study using a far larger database of regional URLs.

7.8 Conclusion on preliminary results

One of the objectives of this part of study was to look at the contribution that web mining for company information might make towards answering the research question:

"Can the use of the internet and the world wide web as an information resource add anything useful to more conventional methods of researching industrial clusters?" Given the difficulties discussed above the answer at this point might reasonably be in the affirmative - but not completely as the whole area needs more work.

Additionally in 2003-4 only about 40% of N.E. regional companies had a published web address although this is rising year on year. It is not unreasonable to expect that sometime in the future this figure would rise to nearer 100%.

Web mining of any sort is a non trivial task. The amount of data generated is large, particularly if links are followed and any one site has many pages to be indexed. Having said this however the basic tools are becoming available, all routes have, at this point, not yet been explored and a watch for other methods or adaptations was continued as the research progressed. A more considered answer at this point therefore might be that the web has great potential for augmenting 'conventional' sources of data and information with the ultimate goal of finding Porter's or Enright's 'hidden' clusters.

As a final note to this section we should not forget cluster identification tools that already exist, the so called conventional methods that were the subject of the previous literature review. This element is important as all the preliminary evidence supports the view that web derived information will at best used as an adjunct to 'conventional' sources of data rather than the primary or only source. Many cluster investigations in the past have involved a combined variety of data sources, techniques and methods and augmentation of this process by web derived information is a logical progression.

Chapter 8. Further Development of a Functioning URL Derived Information Database.

8.1 Introduction

All the work carried out to date has been with the database of 4300 URLs originally obtained from a small call centre and the basis and make up of these URLs was discussed in Section 7.5. At the time it was felt that such a population was sufficiently large to use as a basis for proving the concept of stripping out text in order to gain meaningful information about an individual firm, a group of firms or even potential clusters. As outlined in Section 7.8 in the conclusion on preliminary results, that assertion would appear to be essentially true. However as with many exercises of this nature, the bigger the population the more robust the results and the better the opportunity for examining a greater range of detail. Further, the concept of URL bias as discussed diminishes as the number of URLs approaches the total firm population in a defined region. In other words in a perfect world with a comprehensive set of regional firms each with a known URL then there would be no sectoral bias. The reality however is quite different and obtaining large numbers of accurate URLs is difficult. The term 'accurate' has been used deliberately as although many URLs might be 'valid' as regards there being a functioning website at the address such a web address might not necessarily relate to a functioning economic entity or other organisation necessary for membership of an industrial cluster. This is a key point which is returned to later in this chapter. The decision was therefore taken to look at ways to increase the size of the company URL database. In practice this approach ran in parallel to a number of other lines of research, mostly related to linkages and networks, but for clarity the results of this effort have been written up here as a continuation of the preliminary work using the initial 4300 URLs.

8.2 Basis

There are a number of possibilities to increase the available URLs in a given geographical area and which fall into two broad categories:

1. The first and most obvious one is to acquire more company lists and hope that there are a useful proportion with a given website i.e. a valid URL.

2. Secondly to take the company lists for which no URL field exists and use a search engine on that company and its given details to find its website (if that does exist on the world wide web).

Again it was decided to carry out both approaches in parallel, principally to save time as there was little to be gained by a sequential approach. For clarity the second method is described first as in the event this turned out to be not as successful as anticipated and it is relatively self contained whilst the other method, that of gaining additional data records on firms became something of a work in progress as more company URLs were acquired over the course of time. A detailed description of this continuing process of URL acquisition is shown in Appendix 8. Chronology of URL database build.

8.3 URL Finder

8.3.1 Outline and Function

For most firm data records obtained from commercial or publicly available lists the fields generally include the name of the company, several address lines, the town, county and post code followed by a variety of other data related to contacts and to economic (mostly financial) activity.

Given such data and a search engine such as Google it should be possible to work backwards to find the company website if of course it exists. Trial and error shows that the company name and postcode plus say the post town is usually sufficient to get somewhere near and with further modification a subsequent match to the base URL.

Such a search does require some effort and as with manual searching for keywords described back in the early chapters it would be preferable if the search process for large scale URL acquisition could be automated to run without significant human intervention.

The key however is human intervention as a modest degree of intelligence is required to pick the base URL from the many thousands presented by the average Google search. The situation is further complicated by the presence of URLs which appear first on trade directories and which have reference to company sites but as a subset of their own top level domain. This is not surprising as Trade Directories are organised to have their own URL towards the top of search lists or pay as part of for example Google's AdSense program to be prominent on a returned search page.

One of the early OCW databases from which the 4300 URLs were subsequently extracted contained data records on 13934 companies or in other words there were 9634 without a known URL. This list was potentially a source of further data to enhance the URL database on the assumption that some of the firms did actually have a valid URL now but which for one reason or other did not appear on the original list. In addition it is relatively easy to obtain company lists from the main company data providers for any defined geographical area. The decision was therefore taken to design a program that could input a standard database of company records and automatically interrogate a search engine to return a URL (if it existed) for each individual company searched for.

The flow diagram and underlying strategy for search, validation and output is shown in Appendix 4 Website finder walkthrough.

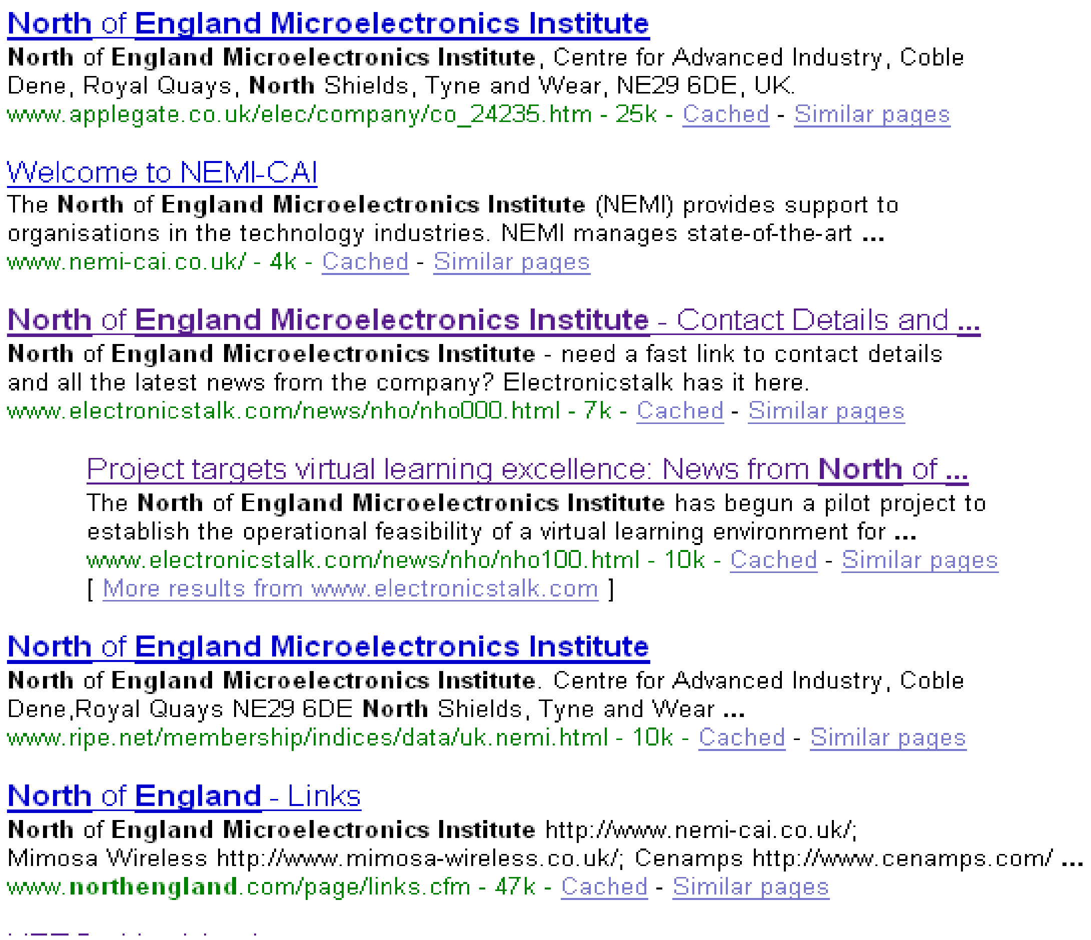
8.3.2 Results and problems

Although the program design philosophy as described in the Appendix is straightforward and logical, actually getting the program to work effectively proved inordinately difficult. The first part which involved stripping off the fields of interest from a database and submitting these to a search engine was accomplished without undue difficulty.

The second part however, that of assessing which of the results was the top level domain name for the company in question proved problematical. As a result the third section which involved comparing data from the determined URL with data from the original list often became irrelevant.

The nub of the problem, that of being unable to mimic intelligent human intervention, caused the program to have long processing times coupled with a very low 'hit' rate. As an example to demonstrate the nature of the problem; the words from the company name and post town for 'North of England Microelectronics Institute north shields' were submitted to Google and a printout of this search is shown in Figure 6 below.

Figure 6 - Search Engine Output Example



North of England Microelectronics Institute
North of **England Microelectronics Institute**, Centre for Advanced Industry, Coble Dene, Royal Quays, **North** Shields, Tyne and Wear, NE29 6DE, UK.
www.applegate.co.uk/elec/company/co_24235.htm - 25k - Cached - Similar pages

Welcome to NEMI-CAI
The **North** of **England Microelectronics Institute** (NEMI) provides support to organisations in the technology industries. NEMI manages state-of-the-art **...**
www.nemi-cai.co.uk/ - 4k - Cached - Similar pages

North of England Microelectronics Institute - Contact Details and ...
North of **England Microelectronics Institute** - need a fast link to contact details and all the latest news from the company? Electronicstalk has it here.
www.electronicstalk.com/news/nho/nho000.html - 7k - Cached - Similar pages

Project targets virtual learning excellence: News from North of ...
The **North** of **England Microelectronics Institute** has begun a pilot project to establish the operational feasibility of a virtual learning environment for **...**
www.electronicstalk.com/news/nho/nho100.html - 10k - Cached - Similar pages
[More results from www.electronicstalk.com]

North of England Microelectronics Institute
North of **England Microelectronics Institute**. Centre for Advanced Industry, Coble Dene,Royal Quays NE29 6DE **North** Shields, Tyne and Wear **...**
www.ripe.net/membership/indices/data/uk.nemi.html - 10k - Cached - Similar pages

North of England - Links
North of **England Microelectronics Institute** http://www.nemi-cai.co.uk/;
Mimosa Wireless http://www.mimosa-wireless.co.uk/; Cenamps http://www.cenamps.com/ **...**
www.**northengland**.com/page/links.cfm - 47k - Cached - Similar pages

As can be seen out of the first 6 results 5 are for organisations whose top level domain name is nothing to do with the domain name sought - ‘nemi-cai.co.uk’. This appears to be a common problem with industrial companies that are captured by on-line trade directories who are organised to get the directory domain name high up the search engine rankings. The URL finder program does not ‘know’ that the first URL found is part of a trade directory and ends up searching the entire directory trying to find a match for the ‘North of England Microelectronics Institute’. If it fails to find one it then picks the next URL found and repeats the process. As even a modest trade directory may contain thousands of company names the prospects for swift convergence are very small. Various improvements to the program such as dealing only with the top level domain, a lookup of the most popular trade directories to eliminate them and numerous other modifications to try and improve the rate of convergence and subsequent validation of URLs found were implemented. These measures did improve the strike rate and returned some useable and accurate URLs against given names. However the process did take a long time and often returned URLs completely unrelated to original company name. It was also noted that for some companies, unlike the automated process for the spidering of web based text, manual methods of lookup and matching could actually be much faster than the URL finder program.

As a result of the above it was concluded that although the problems with the program were probably fixable in the longer term with extensive development work, such a program was a tool to gain additional URLs rather than a key part of the research effort itself and the required time and effort might be better spent on finding additional URLs another way. This sentiment was similar to that engendered by the first ‘Scraper’ program of Section 7.2 in that a decision had to be taken, that further development would be a significant distraction from the primary objective of answering the research question, even though a great deal of work had been invested in program development to the point of use.

8.4 Additional company records

As noted above an obvious way to increase the number of URLs is to access a greater population of companies with a known URL. There are a number of data providers in the marketplace ranging from seemingly cheap CD-ROMs of company lists through to the more established providers such as DNB already referenced. The main ones looked at are discussed below.

EMarket Data

A company called mailing tonic www.mailingtonic.com based in Switzerland supplies URLs under the product name 'Emarket 2005' for a few hundred pounds. The sales literature quotes:

"350,000 e-mail addresses from companies located in UK. Accurate contact data for each company. You can base your selections on a number of criteria. Unlimited exportation! An absolute goldmine for those who intend to canvass the UK market for the lowest possible cost".

It was decided to purchase this data on the basis that such an investment would be low risk in that even if the data was of dubious value it would not be a major loss. The data was downloaded over the web in .csv format and easily loaded into and manipulated within an Excel database. For clarity and understanding of the scope of the data it was sorted by regional county being Northumberland, Tyne and Wear, Durham and Cleveland.

Examination of the data soon revealed a very large number of duplications, incomplete addresses, spelling mistakes and general lack of any quality control. Of particular concern was the very high proportion of URLs that did not seem to relate to companies and a decision had to be taken regarding whether to keep such URLs in the database or not. This could only be done by manually checking the web pages of the given URL and deciding if what was found was a functioning economic entity or other organisation which was likely to be of use to a cluster study. As a result of this evaluation and sorting, in excess of 50 % of the URLs were removed by a somewhat judgemental and hence time consuming process with the following numbers of apparently valid URLs remaining after the cleansing process:

County	**Raw data**	**After cleansing**
Northumberland	1139	678
Tyne and Wear	4964	2050
Durham	2166	1146
Cleveland	1735	811
Total N. E.	10004	4685

Whilst this number of entries appears to be superior in numbers to the 4300 from the OCW database first used, the actual URLs remaining even after cleansing process noted above still gave cause for concern. Primarily an unusually high number seemed to relate to individuals e.g. of the type www.{personal name}.co.uk or others such as www.iloveborome.com

It was clear that the originators of this database, at least as far as URLs were concerned had been merely harvesting domain name libraries without any attempt at checking or any other rudimentary form of quality control. There were other issues in that many companies, organisations or individuals had registered more than one domain name or the same name with different suffixes such as .com, .co.uk, .org etc. Further, many IT companies or web developers had registered domain names on behalf of clients other organisations although it was not always clear when the registrant was acting on behalf of a third party as opposed to merely registering a unique name for themselves. It also appeared from inspection that there may have been many URLs registered by the originator in the hope that a sale could be made of that domain name at some time in the future. This gave a clue as to the age of the data as in recent years the courts have taken a dim view of so called ‘cyber squatting’ or ‘passing off’ of well known brand names or names likely to be confused with such brands and as such the incentive to do so has greatly diminished.

A decision had to be taken on whether to use this data source, as the considerable effort that had gone into cleansing of the records did elicit some useful and clearly valid companies. If it was to be used as part of a wider dataset the downside could be the presence of a significant number of ‘non-economic entity’ websites that added little to any cluster study and which could therefore distort the whole effort, a situation referred to in the early part of the discussion in Section 7.5. It was decided at this point not to discard entirely the URLs apparently available but to look at other data sources and if the URLs from EMarket could add anything after other sources had been explored then these URLs would be considered again. As a cautionary note it had been considered acquiring domain names from a domain name registry on a registered owner address basis. Although this might a first seem an attractive proposition for acquiring every registered domain name with the owner’s postal address in the region the problems noted above do indicate that such an approach would be similarly fraught with difficulties and unlikely to succeed on its own.

DNB Data

Dun and Bradstreet data (now being marketed as DNB data) as previously discussed are one of the leading credit rating agencies and collect data on up to 2M UK businesses. In addition to those companies who apply to be credit rated DNB also collect data on new companies from Companies House registrations. The DNB dataset is the nearest thing to a privately available census of UK business employing more than 10 people.

There appeared to be however one major drawback in the context of this investigation and that is that DNB have not, until recently at least, routinely collected URLs as part of the credit rating data collection process. Nevertheless when approached DNB claimed they were able to match company records with their URLs obtained from one of their associate organisations.

9002 records were thus purchased each with an entry in the URL field. In addition the records came with a subset of the available DNB data fields being:

Keycode
Salutation
Title
Forename
Initials
Surname
Job
Company
Building
Street
Suburb
Town
County
Postcode Telephone
Fax
URL
Site employment
Turnover
SIC (4 digit)

Industry

Activity

Company Type

Head Office

These fields are mentioned as they are useful for other comparative analyses carried out in Chapter 9. In view of the difficulties encountered with the ‘Emarket’ dataset a more cautious approach to DNB data validation was adopted at the outset in that a sample of 100 records was examined in some detail with all the associated URLs being investigated manually. A comparison was then made of the company names and addresses on the DNB list with those found on the same company website. The net result of this short exercise was a perfect match apart from a small number of URLs appearing as unobtainable. It was presumed that the company no longer existed and hence the website had been taken down. A further sort of the 9002 records did find a number of duplicates and some incomplete URLs. The latter were guessed from the company name and again tried manually but the record was deleted if the URL was found invalid. The result of this exercise brought the number of records down to 8518 but at least there was a high degree of confidence in those remaining regarding the validity of the given URL and the quality of the data in the other fields.

NECC Data

During the course of the investigation the project was discussed with the local North East Chamber of Commerce (NECC) with the result that the Chamber agreed to hand over a copy of their member database for use in the study. Because of confidentiality issues only URLs were to be disclosed but for the purposes of this study this was acceptable as we were trying aggregate as many valid regional company URLs as possible. Although full data fields would have been useful as with the DNB data in order to carry out some other comparative analyses it is not necessary when seeking to build an information database founded on keywords derived from company web pages. The number of company URLs obtained from the Chamber was 3648.

Combined Data set

At this point there was available to the project URL data from four main sources being:

OCW	4300
DNB	8518
NECC	3648

In addition there was a possible 4685 from the EMarket exercise. The separate exercise to gain URLs using the URLFinder program was still under development at this point and clearly needing some further work so holding up a main part of the project whilst this problem was solved was deemed not to be a sensible use of time. Although the combined total looks impressive at 21151 records there were a great many duplicates with some companies appearing on all 4 lists. To combine the lists it was decided to leave out EMarket for the reasons discussed in EMarket Data

The 3 basic lists totalling 16493 records were merged and sorted taking care to allow for similar URLs pointing to the same website. Examples of the latter being www.companyname.com and www.companyname.co.uk and www.companyname.net all being one single website and hence a single organisation. All duplicates were thus removed at this time and the net result of this final cleaning and sorting exercise resulted in a database (at this point) of 11580 records.

8.4.1 Additional Data

The combined database referred to above is the one on which the work of the next chapter has been carried out. It has been noted that this combined database is sufficiently large to form a view of the robustness of the methods being developed and used to try and answer the research question and that it is both unrealistic and unnecessary to try and get a regional URL database that is matched in size with all other available company databases.

This is not to say however that opportunities should not be taken to increase the available URL data from whatever source presented itself over the course of the investigation and indeed this did occur.

The program being used and as described in Section 7.4 is such that additional data can easily be added to the combined database as a separate section provided of course that the additional records do not already exist within the Combined Database.

8.4.2 The Influence of Google

It was always recognised that during the period of this project there would be the development of useful tools, the nature of which were difficult to predict at the outset. Further, in a field of study involving the internet and the web with the associated rapid pace of activity such developments would occur with many differing applications but which could be used or adapted at some point to enhance answers to the research question.

The rise of Google has been well known and continues at an unprecedented rate having become large both in terms of its general fiscal worth and its intellectual capability. It is not surprising therefore that Google has come up with a number of publicly available tools that might be of assistance in contributing to the research here. One of these potential areas in particular is that concerned with local search capabilities. It should also be noted at this stage that Yahoo has an almost identical local search capability and in the following description the general methodology used by Google applies for the most part to Yahoo also. The starting point for this local search is accessible (in the case of the United Kingdom site) through www.google.co.uk >> more >> local >> find businesses.

The program is able to find business given a geographical centroid and a business description which can be one word e.g. 'electronics' or 'curry' or any short phrase. Being Google there are of course sponsored advertisements relevant to the initial search displayed in the results. A nice touch is a location map of the found businesses.

The program appears to utilise spidered material from the sites in a manner similar to the Phantom program. The source of the URLs however is interesting in that it draws heavily on a mixture of local and other business directories and when the source URL of a found business is discovered it often but not exclusively refers to some directory rather than the base company URL.

The initial business description must also be used with care as Google makes some assumptions regarding associated meanings. An example is the use of the word 'environmental'. When this word was submitted, in its results Google looked up all businesses associated with 'Environmental Consultants' and also 'Pest and vermin control Services'. Whilst businesses in the latter category mention the initial search term in their descriptors or somewhere on their web site, when looking for evidence of some sort of community of practice for example amongst say environmental consultancy practitioners the presence of large numbers of assorted rat catchers and pigeon eradication firms might distort the picture. The facility does exist however to turn off a category that does not seem particularly helpful and in this respect is very similar to the Phantom program of Section 7.4 used for indexing web derived text. In some respects the use of Google Local as a tool is promising in a number of respects but as with so many proprietary tools and databases built or sold for functions other than academic research its use does require care and a knowledge of the tool's limitations preferably gained through experimentation and proving rather than relying on the suppliers description. Some 99% of Google's income is from the sale of advertising and the use of 'AdSense' technologies, as noted also on page 129, to steer the user towards a product or service of perceived interest derived by tracking web usage is a large part of Google's business model. As noted it is not particularly helpful for academic research in the same way that DNB data is derived for purposes other than academic research. However once these motives and the way in which they function are understood to some degree it is possible to make progress. To contribute to the process here Google Local can be used not only as a search engine for discerning groups of similar firms but also as a source of URLs. This latter process can however be time consuming as already noted many entries appear to quote their source as a directory although it is possible with some effort to track back to the company URL. Google Local Search buys business listing data from commercial data vendors such as telephone companies, although that information may often lack street numbers or other important information, making it difficult for local search to display complete addresses and maps to searchers but it is not known if this is an arrangement between Directory data providers with Google acquiring the basic directory data but then actually spidering the company URL to elicit more company information. Clearly if Google does not quote the directory as a reference and the user can go straight to the firm it does make the use of directories rather superfluous so it may be that after having extracted keywords from the firm URL Google then reverts to the directory as a source for the user. What is known is that Google Local is subject to a number of mistakes[24] in particular with regard to business addresses and it recently (2008) filed a patent aimed at solving this problem.

[24] http://www.seobythesea.com/?p=1022 posted March 2008

The Google patent filing describes how the search engine might try to get more complete information for businesses[25]. In our case however company URLs manually traced back from the directory reference where possible can then be input into the RBKS database to build up the population of URLs feeding the RBKS program. Whilst this may seem a strange duplication of spidering activity, in practice it is surprisingly effective if a little time consuming. There are other difficulties in going down this route again stemming from the reasons already noted with regard to a commercial programme written for commercial ends being used for non-commercial research and some of these are discussed below.

Geographical control in Google Local

Controlling the search area causes a particular difficulty in that the user is asked to input a post code or town as the locus from which to which to search for a particular business type. Google then searches for businesses whose address has on its site that town or postcode that the user is searching for. However Google only searches against those towns that are on the Google map. It also expands the search area when dealing with for example rural businesses which do not have many entries.

An example of this might be searches for ‘Architects’.

If one searches the category ‘Architects’ using the postcode NE1 as the top level location term some 572 are found. This seems an unusually high number of businesses involved with Architects for a modest sized city of 250000 people. However closer inspection reveals however that by the time the entries get down only as far as #14 the postcode has altered to NE4 as Google expands its search area. The default area displayed on the map is relatively small at 1.6 km^2 with 8 entries having their location displayed.

Similarly if one searches for Architects in ‘Newcastle upon Tyne’ Google finds about 522 entries with the word Architects on their website or directory entry but by about entry #50 the address has been (automatically) expanded out of the Newcastle upon Tyne address area. The default area is about 6.25 km^2 with 10 locations being

[25] Local Search Using Address Completion

Invented by Jiang Qian, Assigned to Google. US Patent Application 20080065694. Published March 13, 2008. Filed May 22, 2007 displayed. Locations are noted here as sometimes more than one of the category 'architects' are sufficiently close together to appear as one location. This now becomes more understandable and particularly if the spidered information is throwing up the word 'architects' for sites that are not architectural practices the reasons for such a large count can be understood. At the other end of the geographical scale if a rural hamlet such as Falstone in remote Northumberland is searched there appear to be 4 entries for architects. However on inspection none of these are actually 'architects' located in Falstone and the information gleaned for the website for the Kielder Partnership www.kielder.org for example cites *'The award-winning Kielder Belvedere, designed by London based* ***architects*** *….'* which is a reference to a visual arts structure located quite close to Falstone. Of note is the fact that Google in this case has expanded the search area to over 1400 km^2 in order to elicit the 4 references to architects 'near' the village mentioned. The 4 entries are shown in 2 locations.

In conclusion what seems to be happening is that Google is adjusting the (visual) mapped results to locate the Architects in the exact location e.g. NE1 or Newcastle upon Tyne but that many more references are being brought up as the result of spidering on a wide variety of directory and other sites. In the Phantom based searches Boolean operators are allowed so that more precise terms and combinations of terms can be searched so although Google may at first glance appear to find many businesses of the type requested in a named geographical area it is in fact (a) going outside the area and (b) finding the requested search terms existing anywhere on a company site. Again this is an example of what is primarily advertisement search driven by the need to return some 'useful' results to the user even though in the case quoted above the geography of the search area is vast and one of the references is to an architectural practice almost 400 miles away.

As a final note on architects, the Newcastle Upon Tyne business directory lists 42 businesses under the category 'Architects'.

8.5 Conclusions

This section has described some of the routes taken to try and enhance the URL database. Some, although by no means all of these methods of acquiring URLs through relatively mundane means such as buying data or arranging with a third party to hand over a list of validated URLs were in practice the most useful. By contrast, more sophisticated attempts to automate the process by the use of a custom designed and developed program turned out to be inordinately time consuming and as with the 'Scraper' program of Section 7.2 yielded little and ultimately the decision had to be taken not to proceed with its further use. From the lessons learned it is possible that a focussed effort on solving the problem of automatically finding URLs from name and postcode data could be written and function efficiently, particularly by making an exception to allow for all known directories but the view taken here is that this thesis is investigating ways of answering the research question, not about becoming engaged in a major programming project.

Therefore at the completion of this section the valid URLs numbered 11580 and the subsequent use of these in a research context is discussed in the next chapter. However, as the research progressed any opportunity to add to the URL database was taken and as noted previously the process of expanding the URL database is regarded as a work in progress and the chronology of this procedure is tracked in more detail in Appendix 8. Chronology of URL database build.

Chapter 9. Further Work using the Combined Database

9.1 Program and Data setup

To put the 11580 records in context, they can be but a sample of industry and commerce in the region with the topic of bias being touched on in Section 7.5 in the context of a much smaller database. Again as mentioned the greater the number of records the more confidence one has that some valid results will emerge from processing of the data. In terms of the region however there are 42000 VAT registered companies[26] or an estimated 62000 companies[27] including known others so clearly the numbers used here are a sample currently being 27.5% or 18.6% respectively. As the numbers of URLs indexed increases throughout the life of this research then obviously the proportionate size of these samples will increase. As explained the indexing program can continually add the output from spidered sites as new groups of URLs are acquired but from the point of view of the research we are primarily trying to test the methodology with a significant proportion of the region's business and supporting sites.

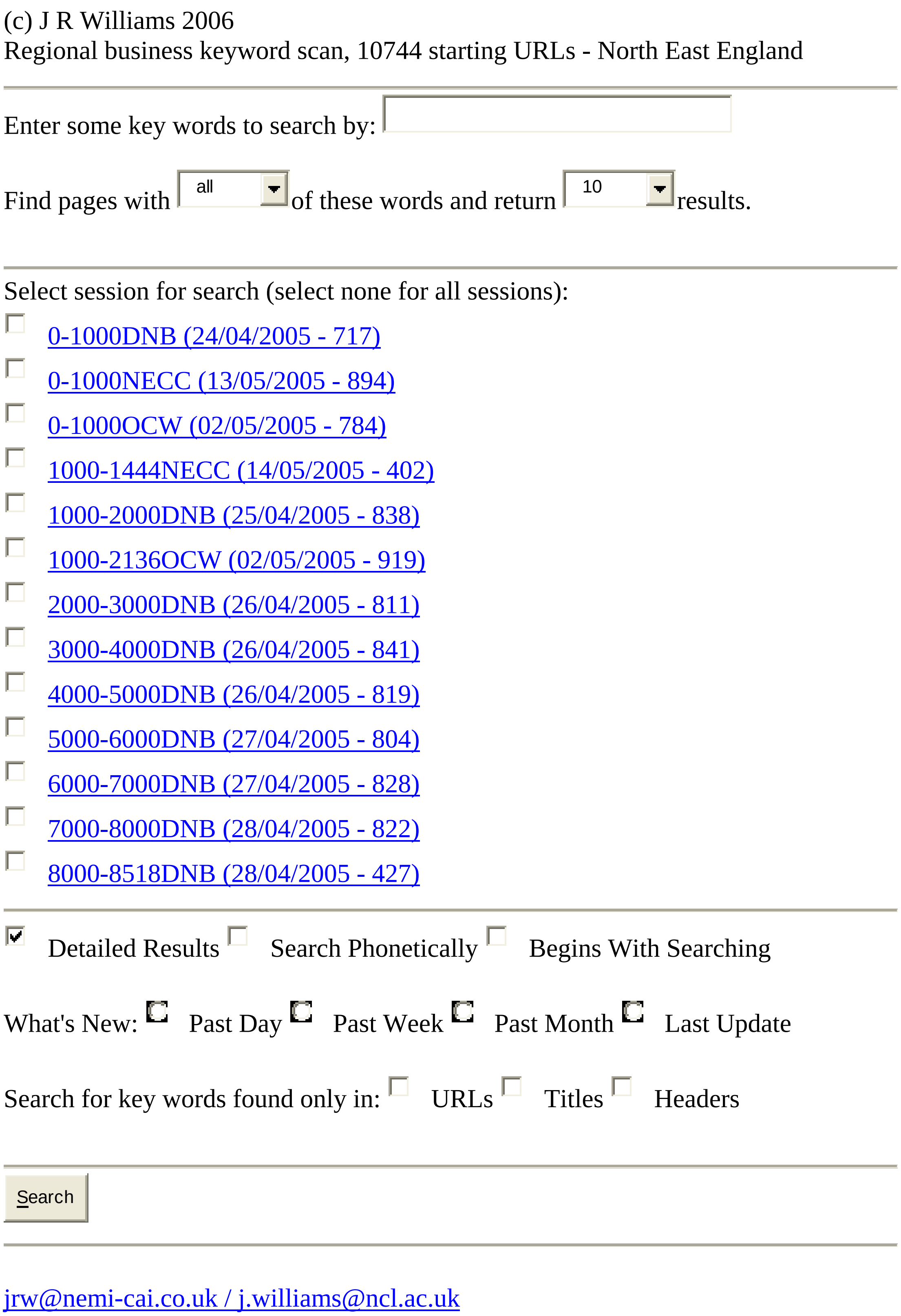
To better reflect the use to which it was being put the Phantom program was renamed the Regional Business Keyword Scan (RBKS), reformatted and then tuned, based upon the many tests carried out with the smaller data sets and reported on in Section 7.7. As a reminder the program parameters for search can be set before spidering commences. To give unlimited depth of search causes the spider to return an additional level of detail generally not justified by the additional time required to undertake such an exercise. For very large sites the timescales can be measured in days and a file size of many Gb. Similarly if the spider is set to 'follow links' it steadfastly follows any internal or third party link on the top level domain being spidered often getting hopelessly stuck on some very large external website mostly unrelated to the original URL. The program therefore was set up for runs to be carried out to depth of 5 but with no links to be searched. This gave a manageable file size and run time but with not too much loss of resolution.

[26] Dept. of Trade and Industry and Office of National Statistics (April 2005), Regional Competitiveness and State of the Regions, Annex 3

[27] One North East, Regional Economic Strategy, 2006

Again to control the program and because of internal programming restrictions each run was limited to 1000 records. If the program did crash or hang on some very large single sites at least the damage was limited to having to restart the 1000 records rather than having such an event occur towards the end of say 10000 records. It should be noted that even with the spider set to a reasonably fast access, each run could take up to 20 hours to complete text acquisition and subsequent indexing of all the words found by the spider. The spidering was carried out initially using a Windows XP based PC with 2Gb of main memory, fast access discs and a dedicated (non-contended) internet connection. However as noted in Section 7.7 the limitation in time is usually with the website being spidered. To speed up the process overall 2 machines were used each with a subset of target URLs with the spidered material being merged when both processes had completed. This simple form of parallel processing thus effectively halved the time taken to spider large numbers of records. The result of all this processing was a set of fully indexed websites based on the 11580 records above. One of the neat advantages of using a URL based system as opposed to a given SIC database is that ‘feedback’ is immediate as when a URL is unobtainable the spider reports back to this effect. Also finding a website from a given URL is not a certainty that a company exists as sometimes the URL hosted by an external provider may outlive the referenced company. If the URL does not exist however either it has changed its URL (unlikely), the website is down (possible) or the company has ceased trading. The latter is the equivalent of the postal ‘gone away’. As a result of this the number of URLs returned as being valid and returning apparently useful text was 10477. However there was a further minor reduction in that a number of sites did not return useful text although they appeared to be valid in all other respects. A manual check revealed that such sites usually had a Flash front end and no metatags, a problem touched on under the earlier evaluation and again reported on in Section 7.7. The starting page for the RBKS program is shown in Figure 7 below and the number of URLs is shown for each section of the total database. The combined total, at this point in the research was 9906.

Figure 7. Example of starting page for the RBKS program



(c) J R Williams 2006
Regional business keyword scan, 10744 starting URLs - North East England

Enter some key words to search by:

Find pages with all of these words and return 10 results.

Select session for search (select none for all sessions):

□ 0-1000DNB (24/04/2005 - 717)
□ 0-1000NECC (13/05/2005 - 894)
□ 0-1000OCW (02/05/2005 - 784)
□ 1000-1444NECC (14/05/2005 - 402)
□ 1000-2000DNB (25/04/2005 - 838)
□ 1000-2136OCW (02/05/2005 - 919)
□ 2000-3000DNB (26/04/2005 - 811)
□ 3000-4000DNB (26/04/2005 - 841)
□ 4000-5000DNB (26/04/2005 - 819)
□ 5000-6000DNB (27/04/2005 - 804)
□ 6000-7000DNB (27/04/2005 - 828)
□ 7000-8000DNB (28/04/2005 - 822)
□ 8000-8518DNB (28/04/2005 - 427)

☑ Detailed Results □ Search Phonetically □ Begins With Searching

What's New: ○ Past Day ○ Past Week ○ Past Month ○ Last Update

Search for key words found only in: □ URLs □ Titles □ Headers

jrw@nemi-cai.co.uk / j.williams@ncl.ac.uk

9.2 Results

The URLs searched by the spider returned text which was placed in a file and then indexed with a link back to the originating URL. The key metrics at this stage of the research were:

File size 60.81 MB

No. of unique keywords 177,432

No. of 'noise' words 231 (not indexed)

We are now in a position whereby a large number of valid company websites have been automatically searched and the text on those sites indexed to the extent that 177432 unique words describe (some of) the industrial activity in the region. What follows in the remainder of this section is an exploration of use of the index as a tool to find out what 'goes on' in the sample of industry and commerce represented by these 177432 words.

9.3 Interpretation

The original question, at least as far as this part of the research is concerned was to determine if the internet could add anything useful to cluster research. Given the power of the internet as an information resource this was probably always going to be true to some degree. After all even manual searching on the internet for additional company information has been shown to be useful. However it was always the aim to find something significant whether that be a completely new method of adding new information or a fast 'industrial strength' search process. This part of the study and the following section therefore tries various schemes to test if the derived company word index does genuinely add anything useful to cluster study methods.

9.4 Comparative work

9.4.1 Keyword Count

A simple test of keyword frequency was set up to compare the keywords elicited by 'normal' SIC methods with keywords found from the Regional Business Keyword Scan (RBKS) using a matched set of records. In the records obtained from DNB there are a number of useful fields for this exercise being 4 digit SIC, 'Main Activity' and 'SIC Text'. Searches were made on keywords representing activity which it is known are hard to find under an SIC based system an example being that of the Motorsport industry.

The results of this simple test are shown in Table 4 and discussed below.

Table 4. Test of keyword frequency for ‘hard to find’ activities.

Word	**Count by DNB 'Main Activity'**	**Count by any DNB 'SIC text'**	**Count by NKS Keyword**
Motorsport	0	0	11
Motor racing	0	0	9
Motorcycle	3	24	38
Automotive	10	8	104
Defence	0	0	48
Military	2	0	25
Offshore	5	0	105
Subsea	1	0	15
Yacht	2	0	13
Sail	1	0	9
boat	6	8	36

9.4.2 Discussion on Test of keyword frequency for ‘hard to find’ activities.

The test was divided into four broad areas each designed to show a comparison of different methods of discerning industrial activity in a geographic region using a known database. The four areas of activity are known to be difficult to deal with as few SIC based categories fit directly. They are broadly; motorsport, defence and offshore engineering and small boat industry. It should be noted that only in Offshore and Defence is the North East regarded as a player and although ‘Offshore Oil and Gas’ is an SIC category such a title often fails to capture many diversified regional firms and other organisations for whom ‘offshore’ is an add-on target market as opposed to a primary one.

The first word chosen for examination here was ‘motorsport’. As described by Henry and Pinch (1999), no SIC at the time existed for motorsport and the activity in the West Midlands which was the subject of their study had to be constructed from an intelligent interpretation of SICs that were likely to be involved in that industry.

Here, for the North East of England there are no entries in the DNB data either under the field for Main Activity or in the SIC text description. By contrast there are 11 companies identified by the RBKS that have included the word ‘motorsport’ on their web site. It should be noted that whilst N.E. England does have a number of specialist engineering design and manufacture firms engaged in this arena the region is not known as a hotbed of motorsport activity so a large incidence of firms involved in motorsport would not be expected. A similar result was obtained with a search on the words ‘motor racing’. By contrast ‘motorcycle’ is a category in Main Activity (3 times) and is well represented under SIC text with 24 incidences in the database. However RBKS returns 38 companies involved with motorcycles. The most marked difference however comes with a search on the word ‘automotive’. Here there are 10 firms under Main Activity, 8 under SIC Text but 105 companies found through the RBKS. An examination of these shows some interesting results. The DNB headings identify firms that are mostly concerned with manufacture of automotive parts, wholesale and retail including retail of automotive fuel. By contrast the RBKS identifies firms involved, not only in the activities just mentioned but also firms and other organisations engaged in a wide variety of supporting roles such as insurance, technical services, specialist recruitment, test equipment, education and training. Here, for the first time we can now see how the RBKS might be used as a tool for cluster study as the plethora of automotive related firms identified from a relatively small dataset (as far as the region is concerned) is much richer that that which can be derived from a single SIC or small group of SICs and in particular it identifies service related firms that otherwise can be hard to discern as their SICs are unrelated to the auto industry.

The next group of words searched relates to another ‘hard to find’ section of industry, that of defence and military equipment suppliers. A search on ‘defence’ showed no results for Main Activity or for SIC Text and two entries and no entries respectively for the word ‘military’. By contrast RBKS identified 48 firms who had the word ‘defence’ on their website, usually under a heading such as ‘markets’ or ‘clients’. A note of caution should be sounded here are as ‘defence’ can relate to other activities other than presumed military options. Examples seen were fencing manufacturers and flood defence systems but these are usually evident from the lines of summary text output by the RBKS front page. As a check 25 entries were found under the word ‘military’ although again there are often unexpected (but not necessarily wholly inappropriate) firms found and an example here was a war games company selling toy soldiers.

On the same theme of more generalised engineering activities being identified or associated with activity in the region that has been suggested as being a Porter type cluster as far back as 1995 [Williams (1998)], a more dramatic example here is that of offshore engineering and support services. A search on the word ‘offshore’ turns up 5 entries in the Main Activity, no entries in SIC Text and 105 under the RBKS. Again there are a few outliers such as ‘offshore banking’ and ‘offshore racing’ but for the most part there is a wide range of firms and other organisations found to be directly or indirectly engaged in all parts of the offshore industry. Although manufacturers are well represented so are a variety of service firms such as lawyers and the Universities.

Drilling down to a subset of ‘offshore’, that of ‘subsea’ activity shows only one entry in Main Activity, none under SIC Text but 15 derived from the RBKS at this stage.

It was noted by observation that many firms identified by the RBKS as being in the ‘defence’ category also regarded ‘offshore’ as one of their markets.

Simple searches on ‘yacht’, ‘sail’ and ‘boat’ similarly show up enough on the RBKS to indicate that some activity exists in the region in these areas. Little activity was discerned by SIC based searches alone. It would be interesting to carry out a similar exercise in areas where a yacht building industry is known to exist such as the South Coast of England or indeed New Zealand [Chetty S. (2004)].

The above points do show that the companies of interest found by keyword search using the RBKS are greater in number, sometimes by an order of magnitude, than those found by either SICs or by searching of SIC text descriptors. However in a cluster research project the researcher would be expected to use a template; that is a guide to additional SICs and hence companies who could be reasonably expected to contribute to a named cluster as in Miller *et al* (2001*)*. Such an approach would of course likely increase the number of appropriate firms that could be members of the cluster under study. The fundamental difference here, in the use of the RBKS is that it removes the guesswork, the so called art when the researcher is called upon to make judgements on the activities of firms on which he may have little knowledge. In other words when using RBKS the companies self select by keywords that the companies themselves have originated (on their website) whereas in the template approach the research methodology or previous researchers have decided, on behalf of a wide range of firms, if they are likely to contribute to the work of a potential cluster. The template approach is only a valid one if the researcher has perfect knowledge of all the activities of all the firms within the database being used – a very unlikely scenario. Given the changing nature of technology and the fast pace of the emergence of innovative products this becomes more unlikely over time. By contrast, as long as the company URLs on the researcher's database are regularly spidered then modifications to company websites are picked up. Many companies have a 'news' section on their website in which all the latest contract wins, new products and anything else of interest to the firm's clients are posted. The spider picking these items up and indexing them is the nearest thing to a real time profile of the developing activities of companies in a given region. A number of specialist market research firms do offer such a service aimed primarily at watching the activities of named competitors. In our case, although it can take some hours of search time, there is no reason why the RBKS should not be set to update the keywords derived from the main URL database on a regular basis and the basic Phantom program has the facility to do this.

9.5 Conclusions on Searches with Combined Database

The immediately preceding Section 9.4 takes as its comparator the set of 8518 URLs from the DNB database described in DNB Data. This is because a full set of DNB fields exists to enable a rudimentary comparison with text information derived from the keyword search program. The general remarks however should equally be relevant to whatever figure the URL totals increase to over the course of the research. It is perhaps worth making a few additional remarks regarding the use of templates to extend the scope of firms that could be engaged in an activity which could be regarded as being part of some supposed cluster. As noted above, in the template approach a known research methodology or work by previous researchers decides on behalf of a wide range of firms if they are likely to contribute to the work of a potential cluster. Such techniques are widely used and are undoubtedly timesaving particularly if the industry and possible cluster being studied say in one state or region is similar to that in another although geographically distant. In this author's view this approach can be misleading and shows a way to the discerning of clusters where no clusters exist. The template approach probably works best with the more traditional industries such as steel making [Slater (2004)] and possibly automotive where the supply chains are well known and identifiable by company although the automotive industry is being subject to changes in its template of contributing sectors. We have long since passed the day when the cost of electronics exceeded the cost of the steel in a car and are heading towards the cost of the software alone being greater than that of the steel. Such technological changes require the template to be kept up to date and in the matter of embedded software for car control systems even this single example would be difficult to implement through an SIC based system.

Another example is offshore engineering as noted in the previous section. There is a perfectly reasonable expectation that firms previously engaged in shipbuilding or marine engineering would sell into offshore related markets if the opportunity arose, on the basis that a floating offshore drilling rig had many commonalties with ships in technology, design, process, materials and manufacture. Although it therefore seems reasonable the practical detail needs care and to include all marine engineering firms and other linked organisations by SIC could easily change a marginal marine engineering cluster into an apparently robust but ultimately erroneous offshore engineering one.

To extend such thinking into an offshore engineering specialisation such as subsea engineering could be even more inaccurate (in template terms). Whilst many of the marine related specialisations such as materials technology and others including fabrication, electrical, hydraulics and control engineering would undoubtedly ‘fit’, others such as lifeboats, marine radar and deck fitments and furniture would have no place in subsea vehicles, a subtlety that would not necessarily be picked up by SIC search methods

By contrast with all the above examples the RBKS picks up a wide range of core activities, specialisations and actual or target markets that firms have put on their website but obviously does not assign them activities that they are not engaged in. Clearly a degree of honesty and clarity on the part of the firm in constructing their website is a pre requisite in the same way that SIC text descriptors are or should be. Generally firms either pay a third party to design and implement their website or invest time and effort in writing their own. It is a reasonable expectation therefore that the result will have been invested with rather more thought than that required to fill in a Government or credit rating agency form describing the firm’s SICs, either as a look-up number or as a few lines of text. It is however acknowledged that firms, particularly smaller ones or early stage companies do use their website as a form of market testing and may not actually be able to do all the things they say they can do on their website.

A final note regarding the corpus of URLs used in the investigation so far. It has been noted that the number of URLs used for the investigations number around 11000. This figure is deliberately imprecise as URLs are coming into being and/or becoming unavailable on a continuing basis, a reflection of the population of firms they are associated with. Although, as has been alluded to previously, the primary interest is to test the efficacy of the method to answer the research question rather than gaining a region wide and comprehensive statement of URLs, the acquisition of further URLs however must be regarded as something of a work in progress as opportunities arise during the course of the research to augment the URL database with additional information. Such additional URLs arise as more firms, particularly smaller ones get a functioning website and the associated URLs are promulgated to various search engines and databases which are then accessed by the author. Sometimes these additional numbers of URLs are relatively small but a continued amalgamation of a number of modest sources would significantly enlarge the combined database. There is of course a problem of bias towards certain groups or sectors as very often Trade Associations are valuable sources of URL databases but obviously reflect the interest of their members which may be closely focussed on some sector or particular aspect of trade or technology.

In view of the above, the search to expand the URL database was continued through the life of the research, largely on an opportunistic basis with the idea being that a single final indexation process would be carried out towards the end of the project.

The final result of this process was that the URL database work tracked in detail in Appendix 8. Chronology of URL database build. was concluded at some 14000 starting URLs of which 12091 yielded what appeared to be valid text information.

Chapter 10. Research into Linkages by the Firm (I)

10.1 Introduction

The flow diagram Figure 1 on page 20 shows the overall plan of investigation with the three fundamental elements that go together to help build an enhanced information database that might be suitable for the study of industrial clusters. The first part regarding text mining of company websites has been the subject of study to this point. The next part of the study looks at how the internet and the WWW might be used in the manner of tracing linkages. This section on company linkages through associations, networks of interest, collaborations and commonalities follows a similar pattern to the previous research on using the internet for text mining except this is in the context of underlying theories of networks and linkages as was outlined in Chapter 5 as they might apply to industrial clusters together with an appraisal of the literature regarding the technologies that might be useful.

In a functioning industrial cluster a measure of its vitality are the traded and untraded interdependencies and the information flows that occur between actors and between actors and the environment. This rich web of interaction has been studied by many authors for almost 40 years [Richter, C. E. (1969,) Czamanski, S. (1971), Lever, W. F. (1972), Dalum, B. (1995), Allen P.M. (1997a, 1997b), Passiante G. Elia V and Massari T. (2003)] on the basis that the greater the flow of ideas and knowledge the greater the innovative output which in turn increases the economic worth of the cluster.

To undertake such a study does however require a significant manual effort to trace these linkages and the difficulties previously discussed with regard to eliciting firm activity information apply at least as much to finding evidence of these linkages in the form of networks. Most networks are established on the basis of multilateral trade or knowledge exchange between companies or organisations and this is often seen by those individual companies as a competitive advantage not to be shared with competitors nor indeed with anyone including researchers. As a result, unless the investigator can be seen as a useful contributor to the network the researchers are unlikely to make progress by asking firms who they buy and sell from, who they deal with and who they gain knowledge from. There are exceptions and most investigators who have succeeded in tracing a significant proportion of firm linkages in a cluster appear to have had some special access either through a trade association or through a sponsoring body or some common interest amongst members in the research outcome. The recent tracing of an emergent subsea cluster in North East England is an example [Andriani P and Siedlock F. (2005)]. Even so, to gauge the full extent of internal and external cluster interaction on a network basis it is usually necessary to visit and question significant individuals in each firm. This is perhaps an indication why most studies of this type tend to concentrate on focussed clusters within a relatively small geographic area.

This whole area of linking of web activity and mapping of the associated topology would, at first glance appear to be fraught with problems in a similar way to that of text mining in that the medium being used i.e. the internet, is full of noise, completely unregulated and not standardised to any universal degree. Similarly the quality of links on individual websites, the reason why firms may be linked to one another and the influence of their environment all vary. This is not to say that there is unlikely to be anything useful emerging from an internet based approach and the next part of this thesis explores the extent to which knowledge of relationships between firms can be thus discerned.

10.2 Network thinking and the research question.

Chapter 5 is a distillation of some of the still evolving studies related to network science and its applications. It is a vast subject touching upon an extensive range of topics in the natural sciences and human activity. In considering the part that networks might play in the study of clusters it is taken as self evident that firms and other related economic entities are connected in some way but it would seem, from the foregoing studies outlined here, that merely regarding firms as simple nodes with links of varying kinds might turn out to be something of a gross oversimplification.

There is also the issue of geography. In the same way that early artisan type clusters started off in a closely defined locale as buyers and suppliers were often constrained by extant modes of transport or the need to be close to each other, so networks became established on a similar basis. With the dramatic rise of modern communication technologies for some types of work such as software production, geographical constraints may be largely removed and Castells in ‘The Rise of the Network Society’ (1994, 2000) discusses how the advent of these communication technologies has affected this balance. In Section 5.3 which discussed network thinking applied to the firm it was noted that much activity was related to multinational and cross border transactions whereas the research question here is looking at activity on a regional scale and within a small region at that. The issue of geography is of less importance where inter-company transactions are being modified in some way by modern communication technologies but here locational constraints must of course be recognised when dealing with networks. Put another way, for some regional firms the majority of their collaborators, information sources and markets may be domiciled outside the region and constraining the study to discern linkages entirely within the region may preclude interesting non-regional connections. This was a problem that was known at the outset and discussed in preliminary chapters with the reasons for adopting a regional approach noted on page 16.

10.3 A look at web linkages.

10.3.1 Preamble.

In Chapter 4 in the context of the World Wide Web, the discussion was on the scope for finding useful descriptive text and for discerning meaning from text. That chapter also discussed a number of fundamental reasons for the difficulties associated with that approach including the almost complete lack of any regulation or standards for the Web and its sheer size and phenomenal growth rate. The chapter concluded that the Web's level of complexity made it impossible to apply techniques from database management and information retrieval in an 'off-the-shelf' fashion and subsequent work on finding useful information was focussed on finding ways around the difficulties.

In the cases of finding linkages however the problems are subtly different. In work done by Chakrabarti *et al* (1999) the researchers undertook an examination of the mining of the link structure of the web in which the notion of identifying the most 'definitive' or 'authoritative' Web pages on a given topic was postulated.

The argument was put forward that as the Web consists not only of pages but of *hyperlinks* that connect one page to another and as this hyperlink structure contains an enormous amount of latent human annotation so it can be extremely valuable for automatically inferring notions of authority. Specifically, the creation of a hyperlink by the author of a Web page represents an implicit type of 'endorsement' of the page being pointed to. Thus by mining the collective judgment contained in the set of such endorsements we might obtain a richer understanding of both the relevance and quality of the Web's contents.

There are many ways that one could try using the link structure of the Web to infer notions of authority and some of these are much more effective than others. This is not surprising as the link structure implies an underlying social structure in the way that pages and links are created and it is through an understanding of this social organisation that we can gain the most leverage. The goal of Chakrabarti and his colleagues (1999) was to design algorithms for mining link information and to develop techniques that took advantage of what was observed about the intrinsic social organisation of the Web. If we are going to examine, for example web based links from one site to another it is worth considering the basis on which these links have been implemented and whether they reflect a firm’s activity by citing buyers and suppliers or if they are more to do with other marketing requirements or social organisation or indeed even the preferences of a particular web site author. A cursory glance at the way web pages of many firms are structured shows up some of the difficult problems we may encounter if we are looking for ways to search for links automatically. First, it is not sufficient to first apply purely text-based methods to collect a large number of potentially relevant pages and then comb this set for the most authoritative ones. As an example, if we were trying to find the main WWW search engines, it would be a serious mistake to restrict our attention to the set of all pages containing the phrase ‘search engines’. Although this set is enormous, it does not contain most of the natural authorities we would like to find (e.g. Google and Yahoo!). Similarly, there is no reason to expect the home pages of Honda or Toyota to contain the term "Japanese automobile manufacturers" or the home pages of Microsoft or Lotus to contain the term "software companies." Authorities are often not particularly self-descriptive; large corporations for instance design their Web pages very carefully to convey a certain look and feel and to project the correct image, often with branding that is not particularly descriptive of the firm’s activities e.g. Orange. Such a goal might be very different from the goal of describing the company. People outside a company frequently create more recognisable (and sometimes better) judgments than the company itself. These considerations indicate some of the difficulties with relying on text as we search for authoritative pages.

Also with particular reference to this section there are difficulties in making use of hyperlink information as well.

While many links represent the type of endorsement as discussed above (e.g. a software engineer whose home page links to Microsoft and Lotus), others are created for reasons that have nothing to do with the conferral of authority. Some links exist purely for navigational purposes ("Click here to return to the main menu") or as paid advertisements ("The vacation of your dreams is only a click away"). It has already been noted that link-based analysis of the Web works best if it is rooted in the social organisation of Web pages. How then can we best understand the way in which authority is conferred on the Web? It was noted above that authoritative pages are often not very self-descriptive; it is also the case that authorities on broad topics frequently don't link directly to one another. It is clear why this should be true for any topic with a commercial or competitive aspect; Google and Yahoo! may all be authorities for the topic ‘search engines’, but they may well have no interest in endorsing one another directly. As will be shown this is a particular and recurrent problem with firms and their competitive instincts. If the major search engines do not explicitly describe themselves as such and they do not link to one another, how can we determine that they are indeed the most authoritative pages for this topic? We could say that they are authorities because a large number of relatively anonymous pages that *are* clearly relevant to ‘search engines’ have links to each of Google and Yahoo. Such pages are a recurring component of the Web: ‘hubs’ that link to a collection of prominent sites on a common topic. These hub pages can appear in a variety of forms, ranging from professionally assembled resource lists on commercial sites to lists of recommended links on individual home pages. Hub pages need not themselves be prominent or even have any links pointing to them at all; their distinguishing feature is that they are potent conferrers of authority on a focused topic. In this way they actually have a role that is dual to that of authorities: a good authority is a page that is pointed to by many good hubs, while a good hub is a page that points to many good authorities [Kleinberg J.L. (1997)].

This mutually reinforcing relationship between hubs and authorities is a theme in the exploration of automated link-based methods for search and potential discovery of thematically cohesive web communities with the potential to lead to additional knowledge regarding the network element of industrial clusters.

10.3.2 Basic Requirements

To consider how the internet, as an information resource might contribute something useful we have to consider the way in which clustered firms (according to most of the definitions in Appendix 3. Clusters: A Variety of Definitions) function and whether electronic communication for example, can be used as a proxy for human to human interaction. In general terms electronic communication leaves a 'trace' of some sort and within limitations might be used as a proxy for evidence of interaction. Such evidence of interaction can be in the form of the who-links-to-whom and who-links-to-me tracking so beloved of search engines when assessing link popularity or more simply by looking at the 'favourite' links on any particular website on the basis that an organisation might show its trading and other useful partners and of course a 'client list' of some description. Clearly firms use email and internet searches, electronic trading and purchasing and a variety of communication facilities to conduct their business. If it were possible to track this spectrum of digital activity it might give an indication of the 'information activity density' associated with groups of firms. Such information on external connectivity might then be applied to appropriate groups of firms and other organisations as found from the earlier section on keyword density. In this case we would be looking for a form of location quotient based on digital interaction instead of employment counts.

There are however a number of very substantive negative issues with the above notion of tracking of digital trades or knowledge exchanges. Primarily, the sheer technical difficulty even without going into the questionable legality of observing such communications. Firms, quite rightly go to great lengths to secure their operations to make sure that their digital activity is secure and that commercial-in-confidence integrity is maintained for electronic as well as for written material.

Secondly there is evidence that face to face communication is important even in the so called 'digital network economy' [Passiante (2003)]. Thus even if, in a perfect world, it were possible and legal to track the sort of electronic traffic noted above it would not necessarily capture all the firm's activity that goes towards the process of networking as a contribution to the process of industrial clustering.

Thus anyone seeking to discern the presence of networks as a contributing factor to cluster development by using direct evidence from wholly digital communication sources is likely to find insurmountable difficulties if attempting to form a complete picture.

However this is not to say that nothing useful can be discerned and a large part of what follows is concerned with examining the limits of information gathering regarding clusters that can be gained from the internet and how this might contribute to the cluster debate, either through an understanding of what can be done directly or as a proxy for reducing the effort required to undertake such studies on a largely manual basis. A further outcome of course might be that with the technology and tools available to us at the present time, not much in the way of useful information can be extracted by web based means that cannot be obtained by existing techniques. Such is the basis of the research question.

10.3.3 Potential tools

In Section 7.1 when considering text mining the form of approach was to consider :

- Existing tools which can be used either stand alone or in conjunction with others
- As above but which require some modification.
- Completely new developments

Here we are dealing with links as opposed to text or 'meaning' but the basic outline proposed is similar although the conclusions may be quite different. What we are trying to do is find links between actors by using the internet and the web to discern these links.

The first and logical route to try therefore is to find any existing tool that enables us to enter a URL and find out who links to that URL or who that URL links to.

10.3.4 Existing Tools

Many of the tools available for discerning linkages are driven, as with text discovery, by the desire for better internet marketing programs. As a result there are many products on sale that aim to improve the visibility of a client URL to the various search engines with the aim of moving a particular company URL to the top of a search engine ranking when a user is seeking information on any particular topic. There is indeed a whole industry engaged in what can be regarded as a continuing battle between the search engines which aim to give a balanced view of the appropriateness of any returned search results to a query and the desire of the URL owner to get their products and services to the top of any search list irrespective of any strong relevance to the original search terms.

Google's current algorithms for choosing the relevance of search results to any user query are a closely guarded secret and anyway they are subject to constant modification as the link popularity industry seeks to discover these algorithms by trial and error and thus propel their own client's URLs to prime position on any search list.

As a result of all this activity, as noted above most of the tools around are available as part of an internet marketing package and a key part of this process are programs to assess link popularity. The term 'link popularity' is used to describe, for any particular URL, the number of links it makes to external sites or more usually and more importantly how many external sites link to the URL in question. The reason for this as implied above is that Google has used such a method for many years in order to assess the relevance and 'importance' of a site to a search query. Such sites, as far as the search engine is concerned are analogous to Chakrabati's hubs and authorities and the idea of 'fitness' in network theory as described by Barabasi (2003) in Preferential Attachment and the Notion of 'Fitness'. More on the technology of this is at shown at http://www.google.com/technology/ with a discourse on the different types of links in http://www.w3.org/TR/html4/struct/links.html

Google's own facility for individual URL tracing of inbound links is available through entering the term "link:myURL" into Google search where 'myURL' is any web address.

There are a number of similar products available now with regular additions seen as search technology advances. Without becoming too detailed they can be regarded as products that seek to assess the current linkages of a given URL. As already noted many of these owe their existence to the prime objective of selling proprietary products and services for improving search results ranking i.e. to elevate the user's URL in any internet search and are thus often followed by a sales pitch on how to achieve this. Three generic examples of this type of link finder are described here.

Firstly a publicly available tool such as Alexa; [http://www.alexa.com/site/devcorner/web_info_services] which shows how an individual URL connects with other URLs . Alexa Internet uses crawling, archiving, categorising, and data mining techniques to build a picture of the Related Links lists for millions of Web URLs. One technique used is to analyze links on the crawled pages to find related sites. The day-to-day use of the Alexa service and related links by all Alexa users also helps build and refine the data. By looking at high-level trends within the millions of URL ‘paths’ created by Alexa users, the relationships between Web sites can be deduced. For example, if many users go directly from site A to site B, the two sites are likely to be related. Next, all the URLs are checked to make sure they are live links. This process removes links that would take the user to pages that don't exist (the so called 404 errors) as well as any links to servers that aren't available to the general Internet population, such as servers that are no longer active or are behind firewalls. Finally, once all of the relationships are established and the links are checked the top Related Links for each URL are automatically chosen by looking at the strength of the relationship between the sites. Alexa Internet recrawls the Web on a regular basis and rebuilds the data to pull in new sites and to refine the relationships between the existing sites. New sites with strong relationships to a site will automatically appear in the Related Links list for that site by displacing any sites with weaker relationships. It should be noted that since the relationships between sites are based on ‘strength’, Related Links lists are not necessarily balanced. Site A may appear in the list for Site B, but Site B may not be in the list for Site A. Generally, this happens when the number of sites with strong relationships is greater than ten, or when sites do not have similar enough content. The two Alexa functions that are probably most useful in the context of finding associations between company sites are ‘Sites Linking In’ and ‘Web Map’. The former action returns the sites linking to a specified web site and Web Map action gives developers access to links-in and links-out information for all text pages in the Alexa crawl. For example, given a URL as an input, the service returns a list of all links-in and links-out to or from that URL. This web map information can be used as inputs to search-engine ranking algorithms such as HITS.[28] The major drawbacks however with Webmap are:

1. The service gives a snapshot of the Web as at January 2005 only and
2. The service was ‘retired’ from June 15th 2007 [http://developer.amazonwebservices.com/connect/ann.jspa?annID=194]

[28] Hypertext Induced Topic Selection (HITS) is an link analysis algorithm that rates Web pages for their authority and hub values. Authority value estimates the value of the content of the page; hub value estimates the value of its links to other pages. These values can be used to rank Web search results. HITS was developed by Jon Kleinberg (1996).

This latter restriction is sufficient to preclude any further work and more importantly checking of any prior work by subsequent researchers.

This does not affect the related SitesLinkingIn action provided through AWIS (Amazon Web Information Services) which gives site-to-site link information. The SitesLinkingIn action is current and will be kept up-to-date, providing a more up to date real-time measure of the relationships between web sites than would have been available with the aging WebMap.

A second site is Link Popularity (www.linkpopularity.com).

This site is billed as ‘The Free Link Site Popularity Service’ and in that respect it does what it says. By entering a URL into a text box Link Popularity will give all related links that it can find from a crawl on Google, MSN and Yahoo! In common with other similar sites there is no restriction (nor is there any control) over the geographical search area. However for single URLs the number of linkages to and from the given site is quite good. It is possible, given enough time, to enter a corpus of regional URLs on an individual basis, record all found linkages to other organisations and try and build up a picture of who links to whom in the region under study.

Another example is Yahoo SiteExplorer (http://sitexplorer.search.yahoo.com) which allows the user to explore all the web pages indexed by Yahoo! Search and to view the most popular pages from any site, build a comprehensive site map and find pages that link to that site or to any page. A particular advantage of Sitexplorer is that the results can be output to a TSV file which can then be read by Excel for further manipulation.

A variation on this would be to take the results from the RBKS in some particular aspect such as ‘offshore’ or ‘defence’ as discussed in Chapter 9 and use the URLs as a seed to find further linkages to the regional firms. Whilst it is likely that such links will be to organisations outside the region there may be regionally based additions. However the RBKS can easily be searched to check if the found link is within the RBKS database i.e. is it regionally based? This theme is returned to later in Section 10.6.2.

There are also a small number of products that find hyperlinks and display them graphically either for search terms or for a single URL and an example of the genre is Grokker[29]. This is a very interesting tool for graphically finding links from and to various sites (but again on a global not a regional basis). It was designed principally as a means of searching federated databases within the enterprise (as indeed was the 'Phantom' program as used in the development of the RBKS) and is particularly suited to portraying the location of information and hence the links between such locations. A neat feature is the use of a clustering engine to show in graphical format the grouping of similar types of information.

Grokker can search the indexes of Wikipedia, Yahoo and Amazon bookstore and display the results either in outline view or the graphical view noted above. However when used for such searches some of the results achieved are often wide of the mark expected and in common with this type of program it is difficult to control on a geographical and hence regional basis.

Another example of the genre is Touchgraph[30] although the basis on which this program works is subtly different from pure link finding programs.

The TouchGraph Google Browser allows exploring of a network of similar pages on Google. Similar pages do not directly represent inbound or outbound hyperlinks. Mutual links contribute to two pages being identified as similar but other factors are also incorporated. For instance frequent third party mentions of the two web pages together will cause them to be listed as similar. To start, the user types relevant keywords or a URL in the search box and presses enter or the Go button. The Go button will display a progress icon whilst data is being loaded, progress being displayed in a status bar.

Searching for a URL will retrieve the top 10 similar pages for that URL and then retrieve the top 10 similar pages for those pages. Searching for a keyword will do a similar thing and retrieve the top 10 web pages matching that keyword, and the top 10 similar pages for each of those pages. By clicking 'expand' the system will then load a further 10 additional similar pages. This process can be continued with 10 loaded each time up to a depth of 30. The website information window at the top left of the screen shows the name of the website, its URL, and a description. This

[29] www.grokker.com

[30] www.touchgraph.com

information comes directly from Google and by clicking the website hyperlink it will launch the website in a new browser window.

The results are shown in a left hand pane which can be sorted by the friendly name of the URL, the URL itself, a number of similar sites found by Google and the grouping of these sites. Perhaps the most appealing feature of Touchgraph however it is its apparent ability to 'cluster' similar sites with a connection from the original source. The graphics output, in a dynamic and customisable form, enables the user to explore connections from a variety of given and found URLs. It must be emphasised that finding similar pages is NOT the same as finding overt or other inbound or outbound links. However it could be argued that when searching for similar sites to a given URL, Touchgraph (and hence Google) is finding sites that perhaps should be linked one to another although in the real world the site's authors would not, for the variety of reasons already discussed, actually implement such overt links on their websites. As with many of these types of program the problem of control is an issue in that it is possible for Touchgraph to come up with the most apparently bizarre connections as a result of following 10 links and then ten links from each of these 10 and so on until the maximum depth of 30 links. Although there may be a certain internal logic to the program's link following mechanism the results can end up with sites far removed from the original subject. As an example, if the URL [www.ndi.org.uk] is used as the seed for Touchgraph it organises its output into some interesting looking groupings using its internal proprietary clustering technology. It should be noted that this is the website of Northern Defence Industries. NDI is a business development company that matches the capability of over 200 companies with the procurement requirements of the global defence and aerospace industry. NDI provides support to international prime contractors, the Ministry of Defence and Tier One System Integrators who require visibility of quality suppliers that can provide collaborative integrated solutions to meet industrial participation and offset obligations. NDI is established in the North East of England and although regionally based does have links with a wide range of organisations on both a global and national basis. As one would expect there are a number of 'similar' sites that would be of interest to anyone looking at NDI's website and indeed examples are the Northwest Aerospace Alliance [www.aerospace.co.uk] and the Defence Manufacturers Association [www.the-dma.org.uk] with 14 and 12 similar 'links' respectively. However the top ranking site, at least as far as similar sites is concerned is *Strategies - Website Design,* *Development & Marketing* [www.strategies.co.uk] with 15 similar links. The reason for this is that the company has undertaken web design for a number of defence related and aerospace organisations such as the Farnborough Air Show [www.farnborough.com], the Society of British Aerospace Companies (SBAC) [www.sbac.co.uk] and Yorkshire Jobs Net [www.yorkshirejobs.net]. It is therefore clear that caution is needed when using this type of programs with frequent reality checks on the results obtained.

On the subject of clustering algorithms applied to either web search for URLs or for links the program 'Clusty' is also worthy of mention. Clusty was started in 2004 when the search software company Vivísimo decided to take its award-winning search technology to the web, Vivísimo itself having been founded in 2000 to tackle the problem of information overload in web search. Rather than focusing just on search engine result ranking, it was realised that grouping search results into topics, or 'clustering', made for better search and discovery. As search became a necessity for web users, Vivísimo developed a service robust enough to handle the variety of information the everyday web user was after. Clusty works by querying several search engines such as MSN, Ask, Gigablast, Wisenut and Open Directory. It can also be customised to use additional sources including a variety of news sites and blogs. It then combines the results, and generates an ordered list based on comparative ranking. This 'metasearch' approach helps raise the 'best' results to the top and push search engine spam to the bottom. What makes Clusty unusual is what happens after the search part is concluded. Instead of delivering millions of search results in one long list, Clusty's clustering algorithm groups similar results together into clusters. Clusters in this context help the user to search results by topic; thus the user can focus more closely in on the desired topic of interest or indeed discover unexpected relationships between items. Generally it obviates the need to trawl down many levels of results from a more conventional search engine but as with the other semi automated programs described above it can throw up apparently oddball results.

Other Forms of Search

There are available products that use social networking techniques to extend the scope of found linkages an example being www.zoominfo.com. ZoomInfo describes itself as 'The search engine for discovering people, companies and relationships'. Although it has a functional search capability, like many sites with a social networking bias it does rely on user generated content, particularly as regards the 'people finder' section. A search for connections to a company only elicited a modest number of references and it is necessary to join and pay a fee in order to test out some of the features such as 'find people at this company' or find competitors to this company'. Many of this type of site are quite North American oriented and the likelihood of finding much useful with regard to a small English region is considered to be low.

There are of course a whole raft of business oriented social networking sites such as LinkedIn and Ecademy to give just two examples. A comprehensive outline, at the time of writing, of Social Networking Applications (SNAs) is given in Appendix 11 Social Networking in Business, based on work by Pollard (2006) but it is fair to say that new specialised groups and associated new forms of SNA appear on a frequent basis.

As these are mostly focussed on people rather than companies they are not considered further here although as was mentioned in 'Other forms of Clustering' on Page 38, the notion of 'connectedness' with respect to people can be shown to be a key parameter in the formation of certain types of cluster.

There are also sophisticated products for discerning patterns of activity within the firm using email tracking software but these products are designed for use wholly within the enterprise. An example of this genre is 'Visible Path'© [www.visiblepath.com]. This is basically a relationship mining engine that scans desktop, messaging and enterprise data to identify relationships and a connection engine which audits 'sales' activity and reveals key relationships in real time. The relationship mining engine uses social network analysis techniques to discern 'important' contacts within a firm and their market relationships. The company literature quotes "In Google, a page is relevant if other sites point to it. In Visible Path, a person is important--we say 'prestigious'--if other prestigious contacts point to them, or if the path to your goal is short,"

A similar product is Metasight at http://www.morphix.com/index.htm

The types of program described in this chapter are not ideally suited to help answer the research question when used singly but when used together or in combination with other methods they can be a useful way of building a picture of relationships between firms and possibly their industrial, trading and support environment.

Whether such linkages and program derived ‘clusters’ are useful proxies for mapping economic dependency between firms is however another matter and this key aspect is discussed later.

To try and summarise the viability of using existing tools for discerning network type activity between firms and relevant organisations there would be appear to be three generic types:

- Links that are embedded in a URL that the site author/s have specifically put on the site as a link of interest to an external site.
- Programs that discern overt links from one URL to another. These usually deal with inbound links to a specific URL in that external sites point to it or reference the site under consideration.
- ‘Traffic’ related links. These are fundamentally different from the preceding two categories above in that the traffic of millions of searches carried out by search engine requests has been mined to determine, for a particular URL, what are the most popular searches and hence these are inferred links. These are referred to as related links and are independent of the other types of links above in that they do not rely on site authors to reference other sites. It might be said that these are links that should be added to a site because activity in the real world shows they are being used as such.

10.4 Existing Tools with Modifications

It was noted above that it might be possible to combine the RBKS and Sitexplorer/Linkpopularity. In practice using Linkpopularity gives the option to search using Google, MSN and Yahoo. As the latter is the basis of Sitexplorer, Linkpopularity effectively incorporates Sitexplorer. By using Linkpopularity to gain knowledge of all inbound links for a particular regional company website a corpus of new sites linking in would then be found. By inspection therefore it would be possible to examine each of these found sites to determine if the site referenced could be regarded as regionally based.

There are difficulties; websites are based in cyberspace, a problem already alluded to and there may be no clear indicator of where a linking site is located. On the other hand there are clues in that sites do give contact addresses and some (but not all) IP addresses can be resolved to give an approximate geolocation [see http://www.geoiptool.com/ for example]. Another method would be to collect all the in-linking URLs found for a particular site and compare them with the database of URLs that is the basis of the RBKS. Although this would not be a definitive view of whether or not a linked URL was in the region (the RBKS database is after all only a sample) it would at least give some indication. In search terms the use of Linkpopularity, even in conjunction with RBKS is very much an 'open loop' system in that the search for links takes place without the researcher being able to influence the search parameters to any great degree.

10.5 New Tools – the Need

Because of the difficulties in controlling Linkpopularity and similar tools on a geographical basis consideration was given to designing a search facility for links that could be controlled. Such a program could be non-trivial and consideration had to be given to the general principle of contributing to answering the research question rather than becoming involved in some major internet search related development. Nevertheless it was felt that there were a few areas that were worthy of investigation to try and obviate or at least mitigate the problems of control of proprietary programs noted above. The development, testing and results of these particular investigations is described in the following sections.

10.5.1 Tracking of look-up history by Cookies

In the early part of the investigation it was noted that the author had acquired a database of URLs from an organisation called One Click Wizard (OCW). The parameters for this data are described in Appendix 8. Chronology of URL database build. To reiterate the basics, OCW was set up to acquire and market company data related to the North East of England and their business model was to give away the data on CD-ROM but have paid-for advertising from vendors which would be visible when the CD was used to search for goods and services. One of the features that the company built into their free CD-ROM was a tracking feature using cookies. .The classical use for cookies is as described in the footnote. However in the case of OCW, as the user of the CD-ROM (when connected to the internet) clicked on various websites of interest a copy of a cookie describing this action was sent back to OCW and collected in a database. The rationale behind this was that OCW would gradually build up a picture of the search activity of the various users of the CD-

ROM. It would also be of use to potential advertisers if it could be demonstrated by such tracking cookies that the population of users were strongly interested in some particular topic.[31]

As has already been mentioned the business model was fatally flawed in that it was ultimately overtaken by the web and the power of search engines but it did seem for a time that there might be an interesting area for research if the tracking cookies could be used to determine who was ‘interested’ in what web sites i.e. a primitive form of linking.

The author acquired a snapshot copy of the cookie messages and analysed them to try and make a decision on whether this was an avenue worth pursuing and if it would yield anything useful to the who-links-to-whom debate. The following observations were summarised from the data:

1. There are 88 companies (browsers) who logged 516 'hits' between 30/10/2001 and 16/12/2001 although the record is not continuous.
2. Total hits 516, Average no. of hits per browser - 6, however two large users have been removed as they had 51% of hits between them and there is a suspicion that they were simply harvesting data. One of these was a recruitment site.
3. 38% only look up their own entry
4. 31% of users look at their own entry first before going on to look at something else
5. 17% do not look at their own business description entry. This could be because they don't have a web site to look at.

[31] A cookie is normally a message given to a Web browser by a Web server. The browser stores the message in a text file. The message is then sent back to the server each time the browser requests a page from the server. The main purpose of cookies is to identify users and possibly prepare customized Web pages for them. When the user enters a Web site using cookies, they may be asked to fill out a form providing such information as a name and interests. This information is packaged into a cookie and sent to the user’s Web browser which stores it for later use. The next time the same Web site is visited, the browser will send the cookie to the Web server. The server can use this information to present the user with custom Web pages. So, for example, instead of seeing just a generic welcome page the user might see a welcome page with their name on

6. 13% look at someone else and then look at their own site or do not seem to have any discernable pattern to their browsing
7. There were 246 hits on the 'site' (48%)
8. There were 205 hits on the 'info' (40%)
9. There were 50 hits on the 'map' (10%)
10. There were 14 hits on the ‘mail facility’ (3%)

The terms ‘site’, ‘info’, ‘map’ and mail refer to self explanatory sections on the CD-ROM or each company entry.

In some respects the information would be analogous to keeping track of a user searching through Yellow Pages except that now the Internet equivalent www.yell.com can easily track the search behaviours from many millions of searches. The only point of difference with the OCW system was that it was known who the users were in some detail as when the CD-ROM was sent to a user (via a postal address) it had within it a line of code that was attached to each tracking cookie sent back to OCW. So whereas a search on Yell.com would be anonymous (apart from their IP address), unless the user chose to disclose their identity, with the OCW system it was known who was looking at what entry on the CD-ROM. Hence the ability to determine #3, 4, 5 and 6 above.

It had been hoped that companies would spend some time looking for suppliers of interest and that some sort of bounded network of such interest would emerge. With a relatively small count of 4300 URLs (although the full database contained details of >13000 companies) this was never going to be more than a sample. The main advantage of having only regional, validated, bona fide companies was that it could obviate the vast amount of irrelevant (from the point of view of normal industrial transactional activity) searching behaviour that might be tracked from a company using the internet generally.

Whilst this may have worked to a degree in an ideal world with a complete list of all regional firms and an active user base searching for firms of interest to their business it was clear from the observations that was not the case and would be unlikely to ever be so.

There was an additional factor to be considered at the time in that cookies had developed something of a bad reputation for acquiring user activity and sometimes personal information which was passed to advertisers and other parties without a user's permission or indeed knowledge. At the time a new version of Windows was about to be released in which the default was for cookies to be set to 'off'.

It was therefore concluded that in this case cookie tracking for this subset of regional industry was unlikely to be of much use to discern accurate linkages that might be useful to help answer the research question and further work on this aspect was not continued.

10.5.2 A Spidering program for Web links

Preamble.

In the preceding discussion on methods and possible tools for discerning networks by linkages a recurring theme is that associated with control of the main parameters of search (or more precisely the lack of control). There are interesting tools around that could be used or adapted for use to some degree but as they have been written primarily for marketing as opposed to research purposes the problem of constraining large scale URL searches for linkages to a nominated geographical area is a very real one at the present time.

As with the decision process for text based search it was therefore decided to look at the possibilities of writing a customised program whereby all the main parameters deemed of importance in a regionally based search regime could be controlled. What follows describes the thinking behind such a program, the development and testing and its use in the context of finding links to and from regionally based organisations.

In the section on searching for web based text in Chapter 7 a spider was employed as an agent to automatically search web based text according to a predefined set of rules specified by the user.

A spider, as already discussed, is basically a program which can be organised to undertake a variety of web related tasks of which text acquisition is just one. In this part of the research we are interested in acquiring information regarding 'links' from websites in the manner of a human observer moving around a chosen site looking for linkages of interest where a 'link' is a reference on a site to a place on another (usually) external site. As discussed with reference to text acquisition, in the matter of large scale scanning of websites it is not practical to undertake more than a few hundred such scans by manual methods because of the time and effort required and because of the scope for human error. A similar situation exists with linkages although the problems here are different and in addition preliminary work has shown that in many cases any particular web site often has no overt links to any other site anyway, not just a regional one. The term ‘overt links’ in this context means that the website author has specifically referenced an external URL and shown this somewhere on the site.

It was decided after some preliminary experimentation that for a link finding spider it should, given an existing database of validated URLs ideally be able to:

- Search individual named sites for valid links to other named sites by following each of these links to a specified depth.
- Locate the subject URLs on a geographical basis or in a form suitable for submission to a mapping program. (UK postcode would probably accurate enough in the first instance, for non-UK then the country of origin would be sufficient).
- Give an idea of the ‘direction’ of a link from a nominated URL to another site even if it could never be possible to assess the magnitude of traffic between sites on the basis of - who talks to me/who do I talk to - as used by search engine ranking systems.

However a problem immediately showed itself when trying out these basic requirements manually and that is that the global nature of the WWW and the unrestricted geography of links was at variance with the desire here to determine regionally based and hence geographically constrained cluster activity.

An example of this is for the site www.maxsi.com as although this site is rich in live links to the company‘s ‘clients’, many of these are outside the region and indeed outside the UK. This is to be expected as Maxsi Ltd is a software and internet specialist with a variety of clients for whom the physical location of Maxsi is of no great consequence. Other examples are www.IBM.com and www.shell.com whose large corporate websites give no single location for their activities. A second more subtle example is the semi-catalogue site www.esources.co.uk who link to ‘partners and affiliates’ on their site. The above examples do at least link to other organisations that the company specifically put down as a link of interest. There is also a third kind of link that many sites show on their pages and that is to tools that might be useful to the reader either in navigation or accessing other multimedia content on the host site. Examples here would be downloads for RealPlayer or Acrobat readers for pdf files, an analogue of the problems noted in spidering for text in The Problem of ‘noise’ on Page 71. In any spidering program looking for links, in the same way that a text grabber does not ‘know’ if the text can be useful knowledge so any link tracing spider does not know which of the types of link noted above it may have traced and thus which is useful in the context of finding cluster type connections, particularly on a regional basis.

10.5.3 Specification, design and program outline

Bearing in mind the above issues and particularly the restrictions regarding global links the outline specification for the design of a bounded link spidering program was designed to incorporate the following requirements:

1. The depth of link finding on the host site should be selectable by the user OR as an alternative the number of pages to be searched by links on any top level domain to be user selectable.
2. The spider should be capable of returning any clear link to a third party site.
3. The output should be such that for a corpus of URLs the links found should be tabulated for easy observation and preferably also for input to graphical or other programs such as for mapping social networks.

It should be noted that at this stage no attempt at filtering links found on a geographical or any other basis was specified. It was felt that the magnitude of the possible problem with global links was not known and that it might be better to just try the spider and see what was returned. However, to try and clarify how such a program might work in practice an outline flow diagram was drawn up of the main elements required to enable the spidering program to function.

Even in outline block format it was clear that a significant investment in time for both learning, coding and testing was going to be required and time spent removing anomalies or requests to undertake searches beyond current technical capabilities would be well spent before coding up. As part of this evaluation process the outline flow diagram in rudimentary form noted below was undertaken manually using feed URLs one at a time. The feed URLs were picked from the DNB dataset on a pseudo-random basis. Thus on any particular website the pages were scanned manually for html based links. As a person and not a spider was looking for these they were not always obvious unless the website author had specifically grouped links under headings such as, for example, ‘Links’ or ‘Client List’ or ‘Partners’ and within these categories there were actual live links as opposed to simple text based names.

This exercise, although time consuming proved salutary in that it appeared that a high proportion of links from regional companies, as already noted, were to organisations outside the region with many other websites having no visible links at all. Attempts to capture non regional links as originally envisaged would not have added much to the knowledge base of intra regional activity but would have resulted in an additional complication to the program with attendant files of miscellaneous links. This requirement was therefore dropped at this stage. As with any web site there were a number of broken links so such a situation had to be catered for in the program. The main change to the initial specification for the spidering program however involved a matrix format of output such that for a defined set of regional URLs as input, a simple ‘who links to whom’ format was adopted. In other words for a regional company citing links to another regional company on the original list this would be detected by the spider and shown in the output in some form of matrix entry.

As a result of these prior tests the outline specification was firmed up and the flow diagram as shown in Appendix 7. Network Analysis Spider - Program Description was drawn. As would be expected this was subject to a whole series of modifications as testing took place but the general outline remained the same.

As with the first URL based text finder described in Section 7.2 attempts to write and obtain a working version of a web spidering program from scratch, whether it is for text or link spidering, is not an easy task because of the presence of so much noise which is so evident when using the internet for this sort of exercise. Theoretically link finding should be simpler than text grabbing because we are trying to discern linkages rather than meaning but in the event the time taken to get the program working at a reasonable efficiency and without frequent crashes turned out to be significant, largely because of the many ways in which third party website authors can embed links. Efficiency here is a measure of the links found by the program divided by the actual number of links within a particular website as determined by inspection.

10.6 Results of tests with Spider program

10.6.1 Preliminary investigations

The regional link finding program as described can output results such that links are portrayed in a simple two dimensional matrix with each URL along the two axes. For any link found '1' is inserted at the intersection with a '0' in all other cases.

Thus for small datasets, by observation it can easily be seen if one URL has any links to another. The program output is in .csv format and can be exported to a variety of other programs for subsequent manipulation. This also includes Excel and although Excel 2003 and earlier versions were easy to use internal programming restrictions limited it to 256 columns. This restriction was removed with Excel 2007 as discussed later.

A testing regime was set up as follows with the aim of using a relatively small number of URLs for each test so that results obtained fully from the spidered URLs could be compared with results checked by manual searching.

The testing incorporated two main themes being

- Quality – are the found connections valid and correct?
- Program checks – metrics related to parameters such as the efficiency of search and results obtained, depth of search, runtime and size of holding files.

To start, the program asks for an input file of URLs. These are in a simple .csv format of the form;

www.fabriam.com
www.maxsi.com
www.semta.org.uk
www.applegate.co.uk
etc.

The program then requests the user to specify the depth to which each page should be searched. This particular aspect was subject to some testing as described later as, in simple terms the deeper one goes the more links one may find. The downside is that the process can take a long time with some very large temporary files being generated and there is no guarantee that the link finder is actually finding anything useful in the process.

During the search process three files are generated including one to hold all the received data and one to hold the temporary output prior to sorting for submission to the final output program. After searching for links has completed the user is asked to specify the file to be submitted to the output program which then sorts the holding file to produce a 2-dimensional .csv file of URLs with links as described above.

Quality checks.

As noted the process was started with a few URL with known links as there is little point in engaging a large dataset of URLs only to find that they cause the program to hang or the system to crash after many hours of processing. By its very nature searching through myriad links using the internet is a time consuming process and even though the program, after five attempts has built-in features to 'jump over' dead links, those not returning any data or other problem links the search process can still be very time consuming. The type of links that can cause the program problems was not fully appreciated at the outset and a number of exceptions had to be added as development of the main program proceeded.

As an example of a number of trials a simple set of the following links similar to those above as input file was used to test the initial trial program version 1.0

www.fabriam.com
www.businesslinktw.co.uk
www.semta.org.uk
www.nmi.org.uk
www.nemi-cai.co.uk
www.applegate.co.uk
www.n-e-life.com/business

The above URLs, by inspection have at least 10 links referenced from/to each other but early versions of the link tracing spider could only find 2. The efficiency of search therefore is 20%. As a result of this experience the structure of the program was again examined. It was always known that Flash sites and those employing frames would give problems to the spider but the number of error messages and poor performance was much more than initially expected. Generally speaking in trying to improve the program quality the only way to do this was to trace the history of a particular search on a specified URL and find out why it stopped or failed to identify a link. Clues to the spider's progress are also given in the temporary files generated which basically keeps the scan history of all sites looked at in a particular session. Although manual inspection and link following was time consuming it did identify classes of common problems which could then be obviated with rewritten code. When undertaking larger runs it was also problematical to have to restart the entire run if the program hung after many hours of processing. To get around this problem which was more prevalent in the testing phase rather than later when the program exhibited better stability and was better able to cope with ‘awkward’ sites a feature was written in to enable a continuation of a program previously stopped. At the start of the program therefore the user was asked if the new program was a continuation of an existing run and if confirmed the new spidered data was added to the existing temporary file holding previous data.

Program Checks

It was noted above that the program has a control feature to allow the user to specify the depth of search for link following. Depth of search for link finding is similar to that for text searching as discussed in Section 4.5 in the context of the use of the Phantom program. The home page would be regarded as the top level with the next level as depth 1, the next click down as depth 2 and so on. With text searching the spider has to index all text on every page and so even after a short penetration depthwise it has started to build up a large database. With link searching there are generally only a few links to find on any site and on most pages there are none. There is of course no general rule for the way websites are constructed but to try and gauge how deep one should search for links a simple test was undertaken initially using just four URLs from the test set as above. It was known by inspection that there were 12 links. The spider was run at various depths ranging from 10 to 400. The results are shown below;

Depth of link searching	Links found
400	12
200	12
100	9
50	9
25	8
10	2

Clearly for this set of URLs there is little point in increasing the depth of search beyond 200.

With small files it is possible to carry out checks quite quickly but they do not really test the spider program from a speed point of view. Therefore a number of test files were assembled to do this but as it was clear that each run was non-trivial in run-time terms the make up of these files was given some thought on the basis that they should be useful and make a contribution to the general objective of finding links amongst regional companies.

One of these files was concerned with looking at the North East chemicals industry. The Regional Development Agency (RDA) has long regarded this sector as strong and the Sainsbury report came to similar conclusions regarding the presence of a chemicals cluster. A quick search of the RBKS showed over a 100 firms with the word ‘chemicals’ on their website. An examination of these showed a few outliers such as cleaning companies but generally they were all companies involved in chemicals production, usage or support services. It was postulated that because the chemicals industry was located to a large degree in the South of the region and that chemicals related industry requires skills and R & D then they might have a relationship with the local universities and colleges. A set of URLs incorporating these chemicals companies and the appropriate further and higher education sector organisations was thus assembled. This resulted in a medium sized test set in which it was felt it would be interesting to observe the degree of web connections (or lack of) and could also be used as test for the time taken to spider the set. After removing some inappropriate sites (see also below) the number of URLs for test was 99. The file was named chemicals99.csv

The results of the spidering for referenced links, time and depth was as follows;

Depth	Time (minutes)	Links found
10	22.48	16
20	48.37	20
30	108.15	18
40	233.63	17
50	275.13	17

From the results of this table it would seem that there is no great advantage to going much deeper than 20. However some of these results are counterintuitive in that one would expect that the links found would rise with increasing depth and then become static as increasing depth of search failed to find anymore or the spider had reached the limit of links followed on the site in question. Here it can be seen that the number of links found actually reduces with increasing depth. To try and find the cause of this a manual search of the URLs was undertaken. Again there appeared to be a proportion of false positives which required amendment to the spider to prevent the problem recurring and other general programming errors which resulted in some found links being overwritten by subsequent ones.

Other problem areas

The issue of false positives had occurred earlier with a different subset of the input file chemicals.csv in that it contained a small number of URLs that appeared to have links to a disproportionate number of other different URLs. When graphed, these companies, of which (www.awakenconsulting.com) was a good example with 15 ‘links’, appeared to be a centre of activity. Closer examination however invariably showed that it was the way the individual ‘node’ websites were constructed that was throwing up links where none existed and the spider again had to be modified to prevent such sites causing errors. Clearly with what was appearing to be a relatively sparse number of regional connections between companies the presence of even a small number of companies with an outstanding number of false positive connections would severely distort the overall results.

A second class of problem was caused by a type of site that had caused similar difficulties with the Phantom searches and that was with catalogue sites. Catalogue sites are a problem because of their sheer size and because the items in their catalogues can be reached with a relatively shallow depth of search. In the RBKS one of the early problems was with a car sales operation which had put their entire range of models, both new and second-hand, provided with a full description only two levels down. The spider in that case started to index every word for every car on the site (although many of them were the same it still kept looking). Although many of these catalogue and other shopping sites are known about (such as www.rscomponents.com with 300,000+ items and which gave such difficulties to the

RBKS when it was in development) others are not always so obvious. Sometimes a URL which once belonged to an industrial organisation but which had been discontinued and bought by an unrelated third party for marketing purposes was then surreptitiously replaced by a reroute to a general shopping site which is not discovered until the spider has been observed trying to find links on the same site for as long as it is left to do so. Such 'holding sites' are quite common and are usually accompanied by a 'URL for sale' notice as well as the aforementioned links to all sorts of advertising including catalogue sites. They are almost impossible to detect by inspection of a list of feed URLs and although seemingly trivial the time they can waste is definitely non-trivial.

When the program was first designed all the data found by searching for links in the manner of the flow diagram shown in Section 10.5.3 was put in a single large temporary file and this was used at completion to feed into the final output program. It was observed that this file could become very large and the problem was particularly acute when attempting, unknowingly or otherwise, to spider such as the catalogue sites noted immediately above. As a result this temporary file could grow in excess of 6 Gb in size. Not surprisingly manipulation of this also degraded system performance and on even a high specification Windows XP machine cpu utilisation was observed to be near 100% for extended periods. The program was therefore again amended such that as each site was completely spidered to the depth specified the results only were written to a separate file and all the temporary data was deleted. Although a single large site still acquired a large amount of data at least if the site was completed the next site had a clean start. The net result of this change was to speed up the process quite considerably when dealing with a corpus of large sites.

10.6.2 Further investigation and initial results

The testing regime described above, whilst making progress on both the quality and the speed of search was not the end of the development process largely because there are some sites whose internal construction causes the spider to stop even though there were by this time a number of built-in mechanisms to try and circumvent this situation. These sites are few but the amount of effort required to find out why the site stopped the spider can be out of all proportion to any useful links found. Sometimes it is better to stop the program, have a look manually for the presence of regional links and note if there are any. At this point it is often found that the site that has caused the trouble is a catalogue or large corporate site anyway having no regional connections whatsoever.

When this part of the research was being designed it was originally thought that the entire URL database would be spidered and that areas of network activity would thus be revealed, largely independent of perceived sectoral influence in the same way that the RBKS can find similar areas of activity in widely differing sectors. This would be in part at least analogous to those social networking sites such as Bebo, MySpace or LinkedIn whereby connections between individuals can be mapped but the difference here is that we would be using companies and organisations instead of individuals. Whilst such an objective might just be possible to some small degree under very, very favourable conditions, early attempts indicated this aim might not yield results on the scale expected but it would in any case be necessary to quantify such indicators. Firstly there was the problem of size. The dataset at this point had grown again to c. 14000 by the acquisition of more URLs on the basis as shown in Appendix 8. Chronology of URL database build. Whilst 14000 URLs on their own might not seem an overly large dataset, for even a modest proportion of this at say 1000 firms this leads to a matrix with 1M cells. For the full URL database of 14000 URLs it would be $(14*10^3)^2 = 1.96*10^7$, a matrix of large size with all the attendant problems of matrix manipulation. Whilst such problems are not insurmountable given enough processing power and memory and enough time there is another difficulty which is largely independent of computing demands. When running larger datasets the chances of running up against sites that cause the spider to stop as outlined above are very much greater. Although during the extensive testing period described above many of the problems causing the spider to stop or ‘hang’ were eventually programmed out. However not all had been removed and although it varied with the websites being spidered it would seem that approximately 1 in 250 sites still gave the spider difficulties even after known catalogue sites have been removed. For the complete dataset of 14000 it would therefore be expected that around 56 sites, anyone of which could stop the program, would be present somewhere during a full scan.

A related problem is one of time. During the test phase the time taken to scan URL lists of various sizes was undertaken. Again the times given are highly dependent upon the individual URLs within the datasets as a single URL can be spidered in as little as a few minutes or as long as several hours.

Table 5 below was constructed by timing a number of runs of varying numbers of URLs. The predominant sector is noted as it is likely that the complexity of an individual website is a function of the time to spider the site for links and it further seems as though the sector has a bearing on the complexity. For example the first line of the table refers to 77 URLs in a category principally related to public relations. The sites by inspection are often rich in a variety of text, pictures, audio and video clips all of which can take a considerable time to spider. The next line refers to 99 organisations involved in chemicals where the sites are more pragmatic and straightforward in terms of a spider searching for links to the specified depth (although there might be a site where a very large chemicals product list is available on-line). Similar remarks can be made regarding the 'marine' related searches. The 'defence' sector is slightly different in that although there are similarities in web design philosophy, as with marine the presence of some very large sites e.g. www.boeing.com and www.baesystems.com slow the process down considerably. It could be argued that as these latter two are not regional firms they should not be in the list and are unlikely to show any connections to regional firms anyway which is the whole point of spidering for links here. The net result is that the table and accompanying graphs show a trend rather than a definitive relationship between the number of URLs in a dataset and the time taken to spider those URLs for regional linkages to a depth of 50 pages. It should also be noted that a further variable bearing on the time taken is depth of search but to reduce the variables this was set to 50 pages for all searches shown below.

The trend from the 5 test cases was projected on the basis of (a) a linear projection and (b) a power law projection. It is not known what the correct form of projection should be although Barabasi (2003) suggests that power laws play a part in such networks. An obvious way to gain a better understanding of which projection would be best would be to run more tests and best fit a polynomial line to the found points. However as the collective runtime required to construct even this modest table was almost 80 hours it was felt that such an exercise would not be worthwhile as the whole point of doing this was to assess the times likely to be required to spider a large regional dataset.

Table 5. Time to spider URL data by predominant sector

Time (hours)	**No. of URLs**	**Predominant sector**
5.66	99	Chemicals
10.16	200	Marine
10.2	77	Public relations
23.84	346	Offshore
29.28	290	Defence

Using the two projections noted above (linear and power law) with data from Table 5 the two graphs shown below in Figure 8 and Figure 9 indicate the projected time to spider a dataset of 1000 URLs. As already noted this would result in a matrix of 1M cells.

Figure 8. URL links vs. Time (Linear projection)

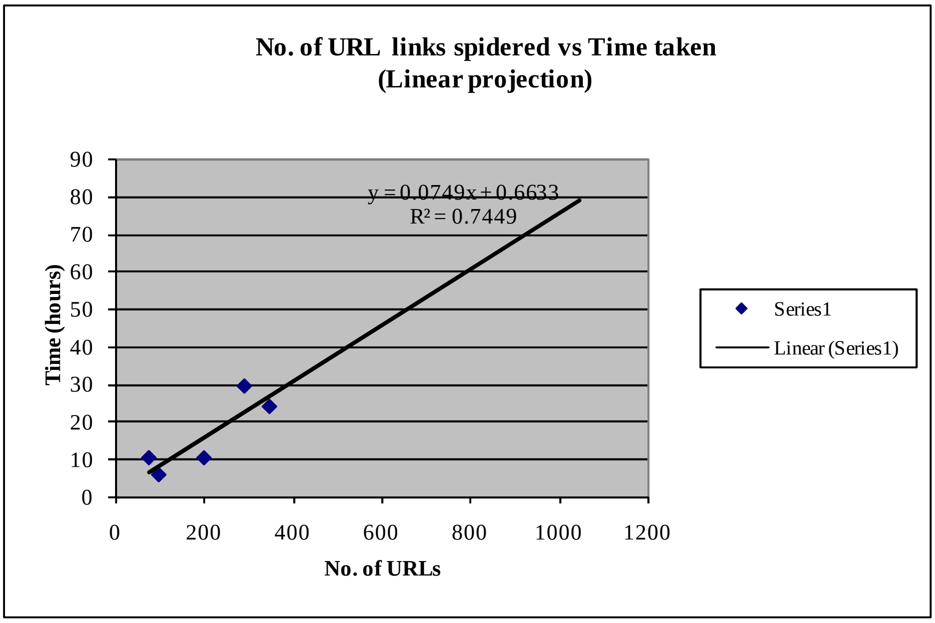
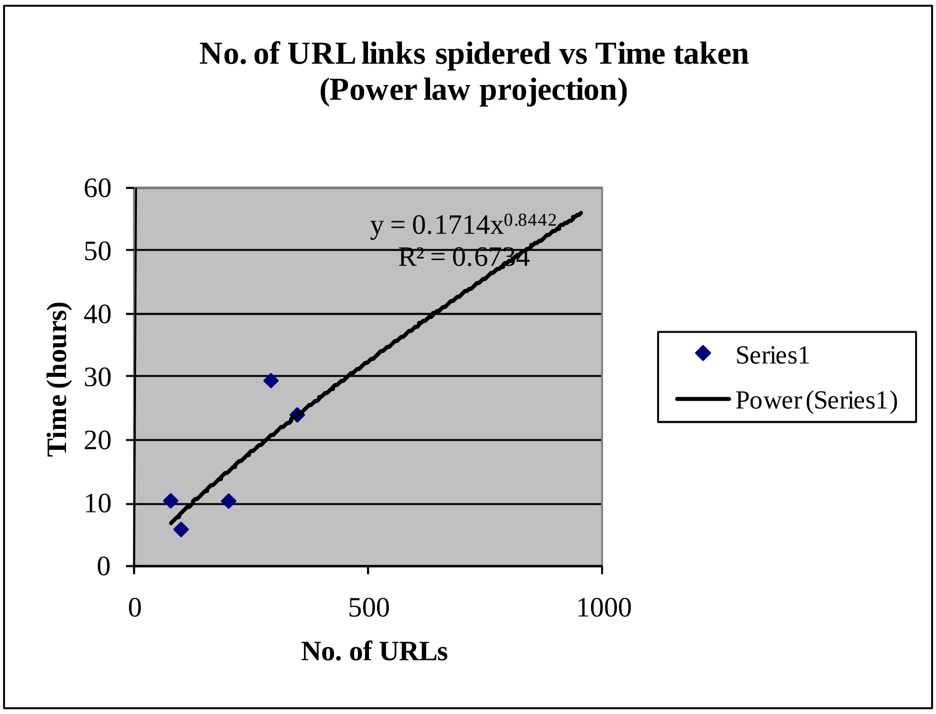


Figure 9. URL links vs. Time (Power law projection)



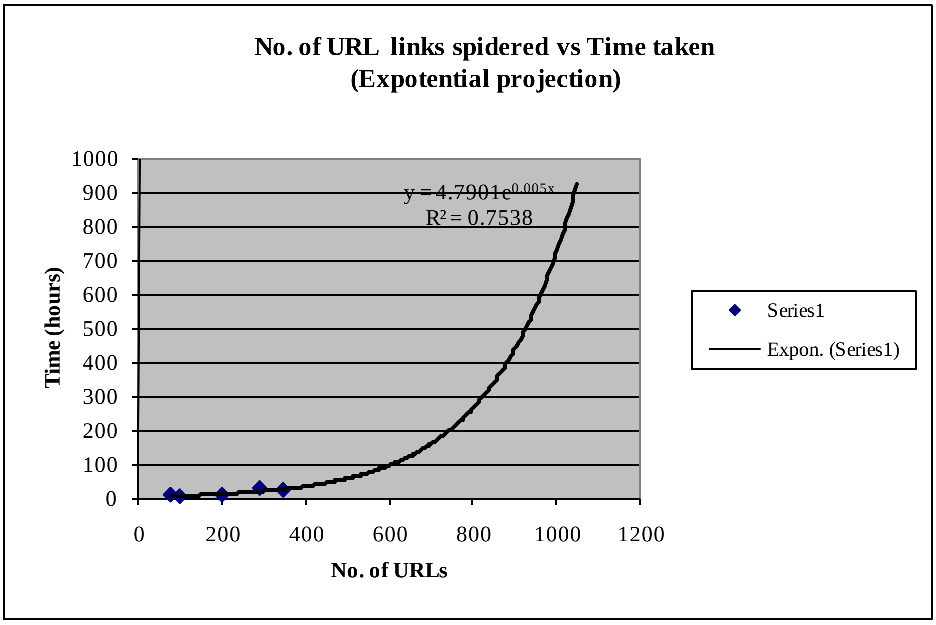
The projected times to spider 1000 links are summarised in Table 6 below.

Table 6. Projected time to spider 1000 regional Links

Projection	**Time (hours)**
Linear	75.6
Power Law	58.4

A further projection using an exponential fit was also looked at on the basis that as the output matrix became larger it was more difficult for the computer to manipulate it and the whole process slowed. In the early days of spider development this was certainly the case as all results were held in the temporary file shown in the flow diagram in Appendix 7. Network Analysis Spider - Program Description and this file could become very large (>6 Gb). However when the program was redesigned to empty this holding file after each successful individual URL spidering the problem largely went away. The final file leading to the output could be large of itself but the computer could easily cope with this and even the largest group of URLs spidered only took about 15 minutes to produce the final output matrix after the spider had stopped. In some respects therefore the exponential projection of Figure 10 was of historical interest rather than practical utility, particularly as the projected time for 1000 URLs would seem to be >800 hours using such a projection.

Figure 10 - URL links vs. Time (Power law projection)



It can be seen that the R^2 value for all three projections is not particularly strong, largely on account of the fact that there are only 5 points to work with and each of these points carry non-trivial error bands for the reasons already noted above.

10.7 Conclusions

This chapter has been a first assessment of the requirements necessary to undertake a study of web linkages. From a view of the basic requirements to do this, it has followed a road map similar to that for text mining in that a number of routes using existing tools were examined. By tools we mean a variety of programs, specific methodologies or some means of gaining the type of knowledge required to explore linkages and potential networks having used web derived data and information. The broad conclusion of this first part of research into firm linkages was that, as with text mining technologies there was no one tool that was ideally suited for the task and was necessary to develop a custom program. The development of this spidering program turned out to require a significant investment in terms of development time and all the problems of dealing with the internet for spidering text were present to some degree for the spidering of links. As the link finding program was developed it was deployed in a number of areas to test both its efficiency at finding lionks and also the limits of practical use. The basic conclusions from this exercise are clear if somewhat simple. To undertake link spidering on a closed basis is very time consuming if the size of the matrix generated is much above 500 x 500. This then leads to the conclusion that whole-region spidering is probably not a viable way forward, at least with the single PC based approach used and described here and that a more focussed approach to the identification of candidate URLs that might be linked should be considered. This is the thrust of the next chapter.

A comprehensive conclusion on both this chapter and the one that follows regarding all the research on linkages and associated networks is given at the end of Chapter 11.

Chapter 11. Research into Linkages by the Firm (II)

11.1 Examination of specialised groups of firms

Introduction.

The work of the preceding chapter concluded that large scale spidering for connecting links was not a practical measure but in some respects it was only tried to try and short-cut the route to identifying connections between groups of firms by the apparently straightforward but in retrospect perhaps naïve precept of starting with all firms on the database and looking to see what emerged.

There is however another route to doing a similar exercise and this is by using the RBKS to identify firms with related activity and then using the link spider to see if they have within that group links at a greater frequency than might be expected from a general company population of the same size.

In running the RBKS it is entirely possible to look for groups of similar firms in the same or related area of business by the simple expedient of searching on a chosen word. An example might be such as ‘environment’. The RBKS has a ‘starts with’ feature that can be enabled such that the word ‘environmental’ appearing on the keyword list would be picked up. As all 177,432 words in the RBKS database are fully indexed and stored locally such searches are very rapid and it is feasible to go through a list of all the main activities that the user feels would be appropriate for the region.

However it is known from previous work by Williams (1998) and by Andriani and Siedlock (2006) that some rudimentary clusters do or at least did exist although some of them are quite small in terms of their ‘membership’ as might be expected as we are dealing with a relatively small English region. Some of these possible ‘clusters’ have already been spidered for links in order to evaluate the run times as shown in

Table 5 and the obvious course of action therefore was to utilise these groups already spidered together with new ones.

It was decided to undertake additional spidering on the following segments of industry:

Electronics – With an SIC based search the firms that are found tend to be manufacturers whereas electronics companies appear ubiquitous when found using the RBKS as it will find any firm with the words ‘electronic’ or ‘electronics’ on their website. Thus a national firm of bakers based in the region does in fact employ electronics technicians to maintain its automated cooking equipment and such expertise can be regarded as part of an electronics milieu. The author’s company is also involved with the electronics industry in the region and has some additional knowledge regarding firms and support organisations that can be used for comparison purposes.

Architects – for the reasons touched on above this was felt to be a good candidate professional community for discerning linkages.

Environment – A search of words starting with ‘environment’ (which would include ‘environmental’) elicited some 775 URLs. At 5.5% of the total URL population being looked at this was a surprisingly high proportion, far higher than any of the other ‘traditional’ clusters for the region and as such is worthy of closer examination. Although 775 URLs are some way beyond the 500 initially seen as a practical limit it was decided to persevere with these and as a first step remove any URLs that were outside the remit for a functioning cluster or community of practice e.g. some companies have a web page entitled ‘Our Environmental Policy’. By inspection this would seem to have more to do with a corporate social responsibility agenda and the desire to promote the firm’s environmental credentials rather than an indication of direct engagement in the environmental industries business.

In addition the RBKS has a facility that ranks the found URL for relevance to the initial keyword(s) search. To further reduce the number of companies, any with a ranking of less than 5% on this rating basis were discarded although even at 5% there was often some relevance. By this process the original 775 URLs were thus reduced to 386.

It should be noted that this is still a significant proportion of the combined database and may be a reflection of the general interest in environmental issues currently prevalent. It may also be that many of the region’s stock of general engineering businesses and technical consultancies see the environment as a growth area and have started to develop and note their capabilities in this area on their websites. This is not to immediately say that there is an embryonic cluster in environmental science and engineering but it may be so and one of the strengths of the RBKS is that it can identify such changes much faster than agency based publications using SIC fields which may contain records several years old when collated and published.

It may also be an indication of another advantage of the RBKS and this is the ability to discern firms in multiple clusters of whatever form or maturity. Thus the same engineering firm can be seen in separate searches involving 'marine', 'offshore', 'nuclear', 'precision engineering' and 'environment'. This is a theme that is returned to later in the thesis as more in-depth methods of tracing links were developed.

Defence – This was a grouping that came in for a lot of attention when doing key word searches and has also been mentioned above in the context of testing and some of the apparently anomalous connections seen between firms. It should also be noted that some of the companies with 'defence' related activities are located outside the original North East remit because they came off the NDI website which has Members from outside the region (although sometimes they have a regional office even though the corporate website has a corporate, or no address at all).

Offshore – as with Defence' this group of organisations had been examined with the RKBS and it seemed logical to look for web derived connections. Again as with Defence the Members list for the Northern Offshore Federation (NOF) was used at one point to gain additional URLs. NOF has in recent times re-branded itself as NOF Energy and is actively seeking new Members from outside the region. Their list, as far as the RBKS is concerned does contain some out-of-region companies.

Marine – In some respects marine activity in the North East of England was the precursor of the Offshore segment and was born out of the significant shipbuilding and marine activity that existed in the area from the 20th century finally dying out with the closure of most of the yards during the late 1980's and early 90s. Only Swan Hunter Shipbuilders survived as a mixed (Merchant and Naval) yard finally succumbing in 2001. The significant marine supply chain and the untraded interdependencies had by this time, as much as it was able, diversified into marine markets overseas and into other related areas.

Nevertheless a search on the RBKS does elicit a large number of firms who regard themselves as being in the marine or marine engineering business and these were spidered for links to see if there was any connecting or collaborative activity that might be an indication of some sort of cluster.

Public Relations – The final group study was something completely different. The North East is not known as a hotbed of PR activity most of which, for larger organisations as least, seems to take place in the capital. However there was some PR activity with the promotion as such coming not surprisingly from the PR companies themselves. It was felt that in a relatively small cohort of firms everyone would know each other and who their competitors where and might cite connected service firms such as lawyers, accountants, graphic designers, web hosting firms, advertisers and the like.

11.1.1 Initial Results of Link Spidering

The time taken to spider links to the depth already discussed was significant and in some respects the results were both disappointing but not entirely unexpected. The reasons for this are discussed later in the Conclusions to this chapter.

In order to portray the degree of linking found for the various segments chosen it is possible to just count the number of links as a proportion of the number of nodes (websites) or, more usefully, to use a network visualisation program to portray the linkages as a network. By inspection the connections can then be seen to be either noticeable, rather sparse or in one unexpected case non-existent.

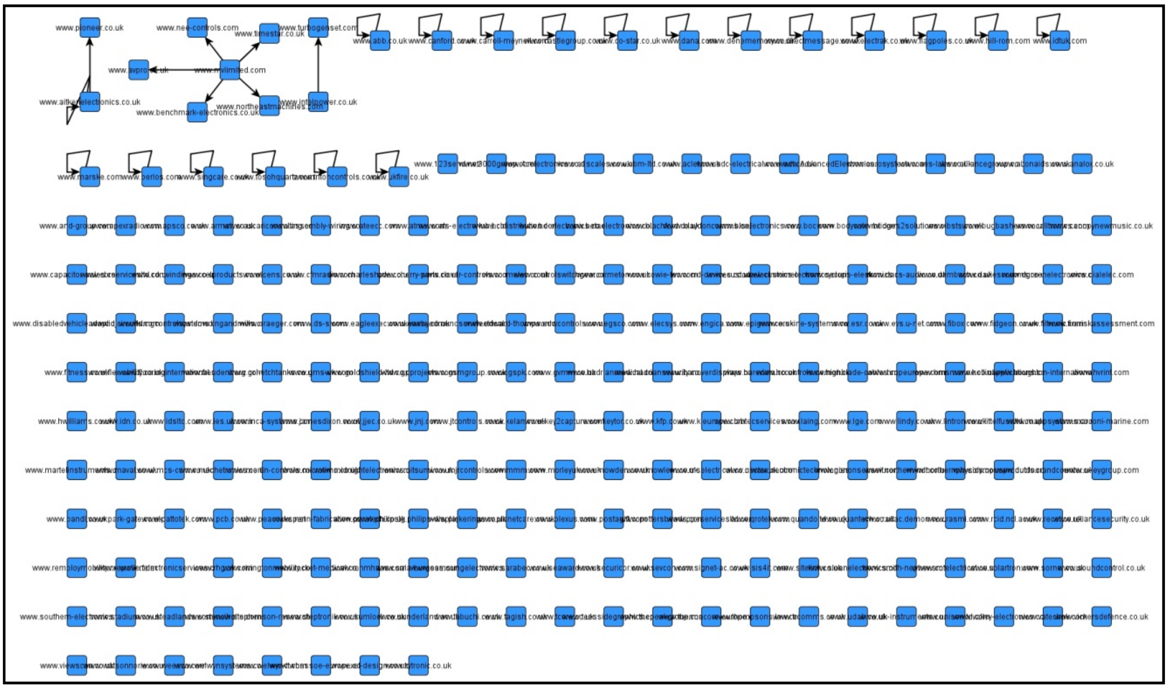
There are a many social network programs that will do this fairly simple task at the 2-D level and for this preliminary look at the results VISONE (www.visone.info) was used. The link finding program outputs files in .csv format which are initially read into MS Excel for a preliminary count of the connections. VISONE accepts as input .csv files from MS Excel and as a further advantage when using MS Excel 2007 the limit for columns is now 1M as opposed to MS Excel 2003 which had a column limit of 256 and which effectively precluded its use for any significant output files greater than this limit.

For the groups spidered so far diagrams and brief comment are shown below.

Electronics - Figure 11 and Figure 12

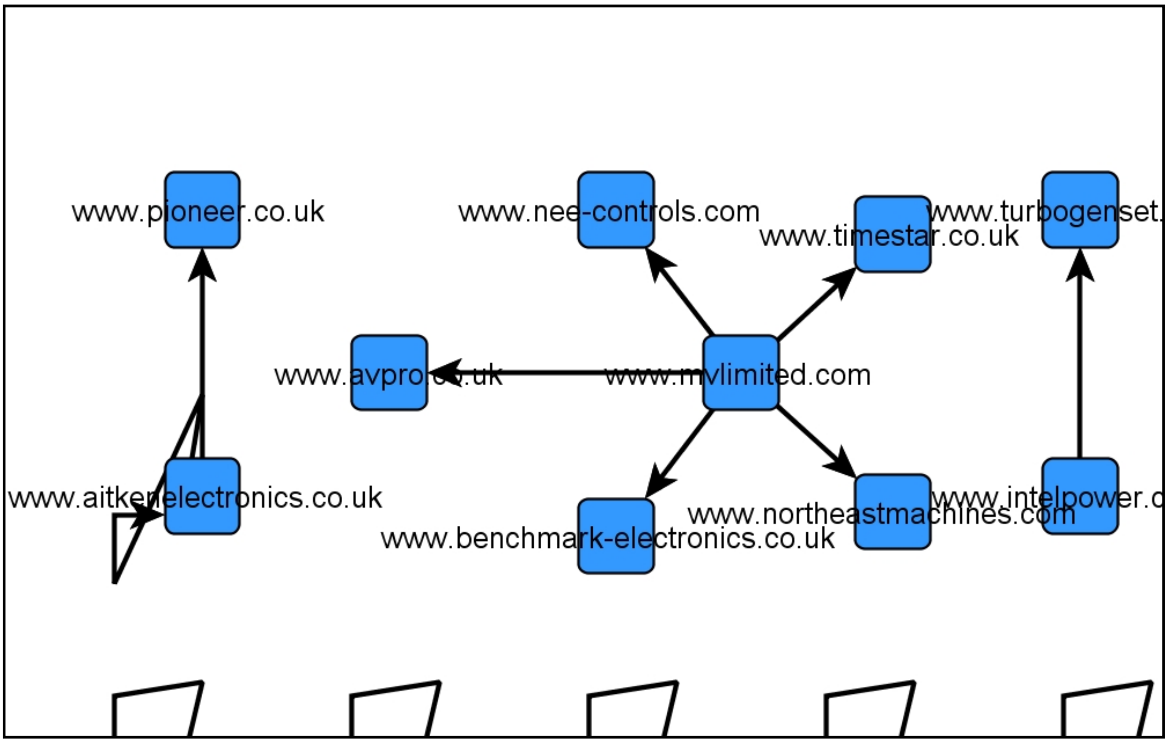
Out of 248 sites spidered there were 26 connections to other sites but of these 18 referenced their own site so there were only 8 true external links. Of these www.mvlimited.com referenced 5 others.

Figure 11 – Electronics (all)



.

Figure 12 – Electronics (detail)



Architects.

Out of 254 firms with the word starting with ‘Architect’ somewhere on their website not a single one had any embedded or clickable links to any other site on the list. This is a surprising result as one would have thought that Architects would be a good candidate community of professional practice and there would be many connections amongst the supply chain. Clearly this result needs some examination as regards the reasons.

No diagram has been drawn as there are no connections to portray.

Environment - Figure 13 and Figure 14

This cohort was based upon 368 firms of whom 102 displayed some linkages. 59 were linked to some other part of their own site leaving external 43 linkages. The most notable was to www.defra.gov.uk indicating the prominent part played by a government agency and which in cluster terms would be an ’non-traded dependency’.

Figure 13 – Environment (all)

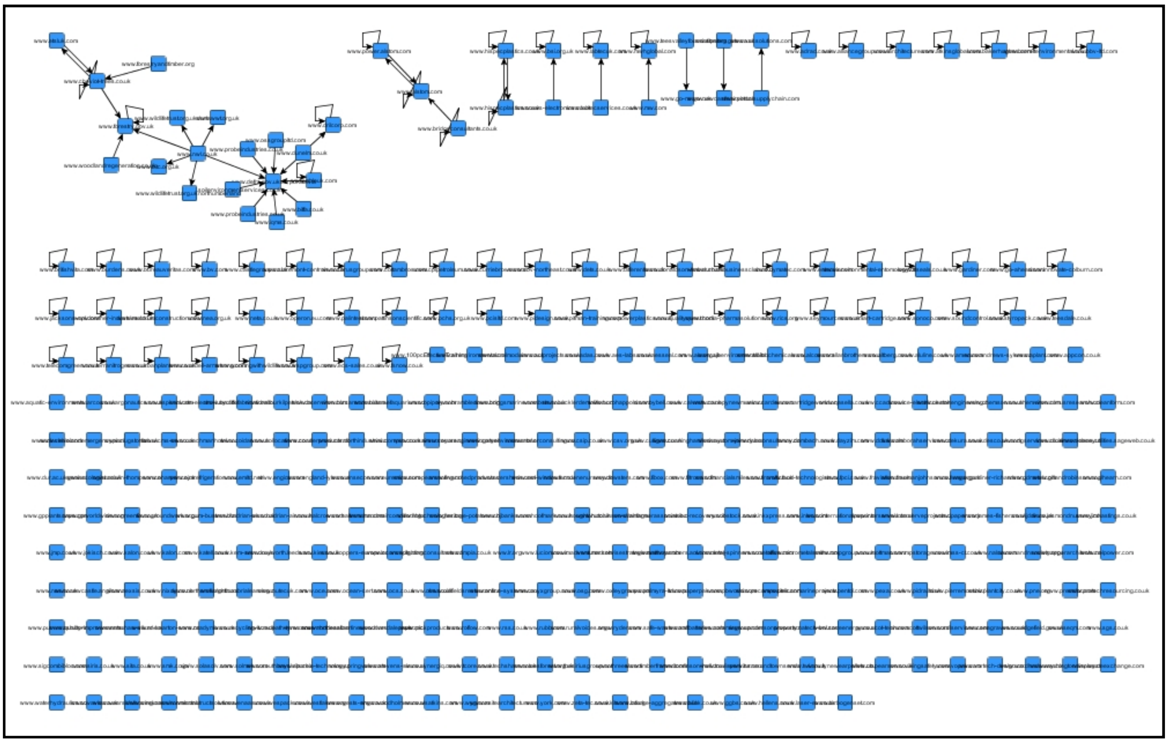
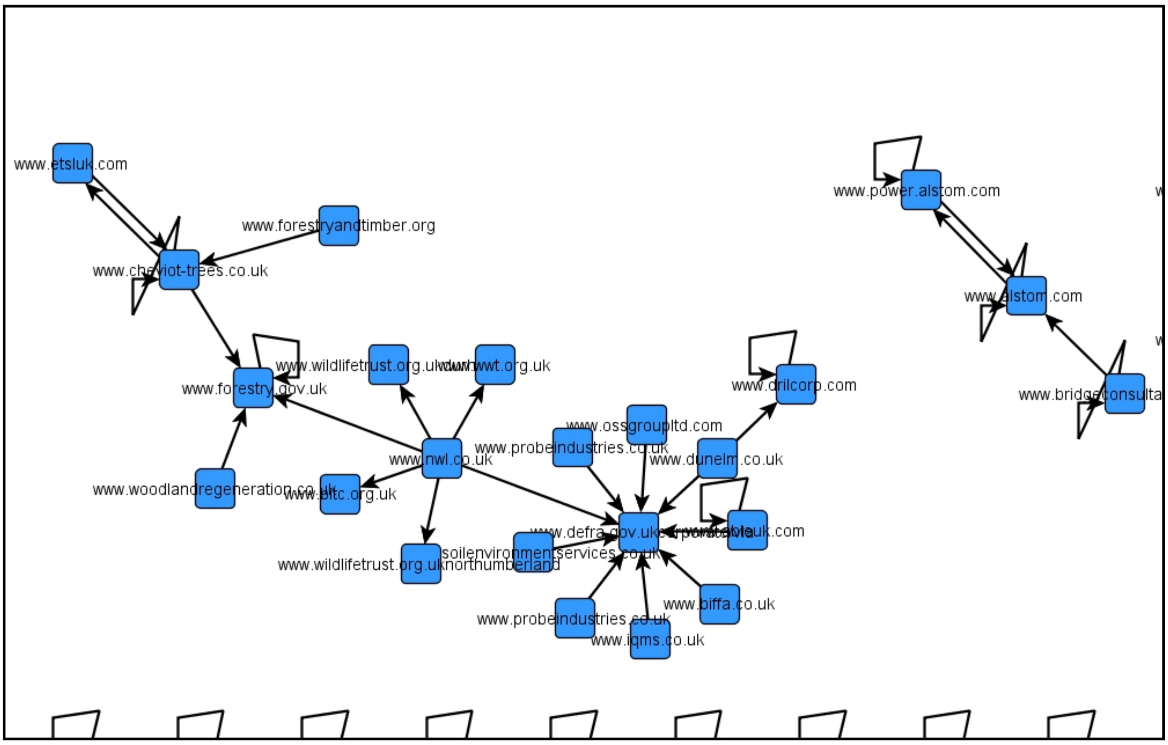

www.forestryandtimber.org
www.probeindustries.co.uk
www.probeindustries.co.uk

Figure 14 – Environment (detail)



Defence - Figure 15 and Figure 16

This is an interesting grouping in that it is starting to show at least some semblance of internal connectivity. There are 92 connections out of a population of 285 firms and of which 43 have circular connections to their own site leaving 49 other connections. Of these the notable nodes are www.ndi.org.uk and www.argonautics.co.uk both of which are trade association type organisations set up to promote what they regard as a cluster of some sort. In the case of NDI this is referenced by a number of inbound firms and its role has already been mentioned previously in Section 10.3.4 and on page 163. Argonautics are a small group of marine companies who work together on combined contract bids and also a joint lobbying basis and have a clear interest in the defence sector through Naval ship design and support contracts.

Also Defence Solutions Ltd (www.defence-solutions.co.uk) are a good example of the case of a consultancy firm citing links to clients it has supported.

Figure 15 - Defence (all)

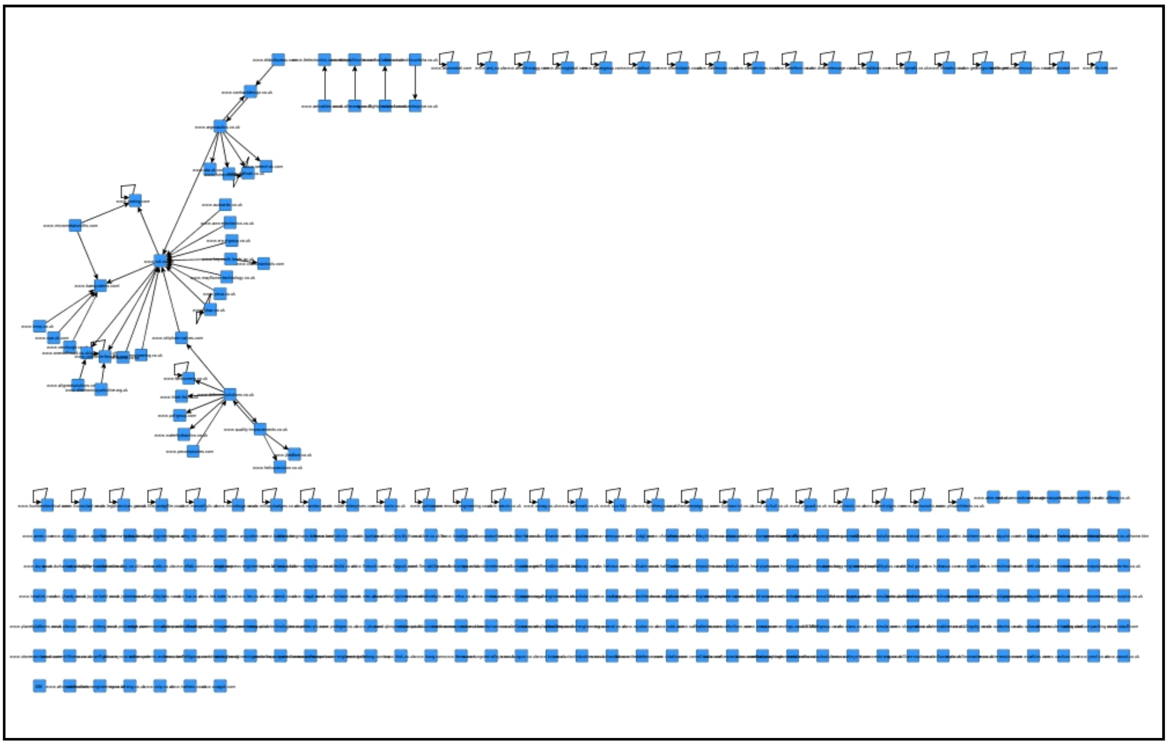
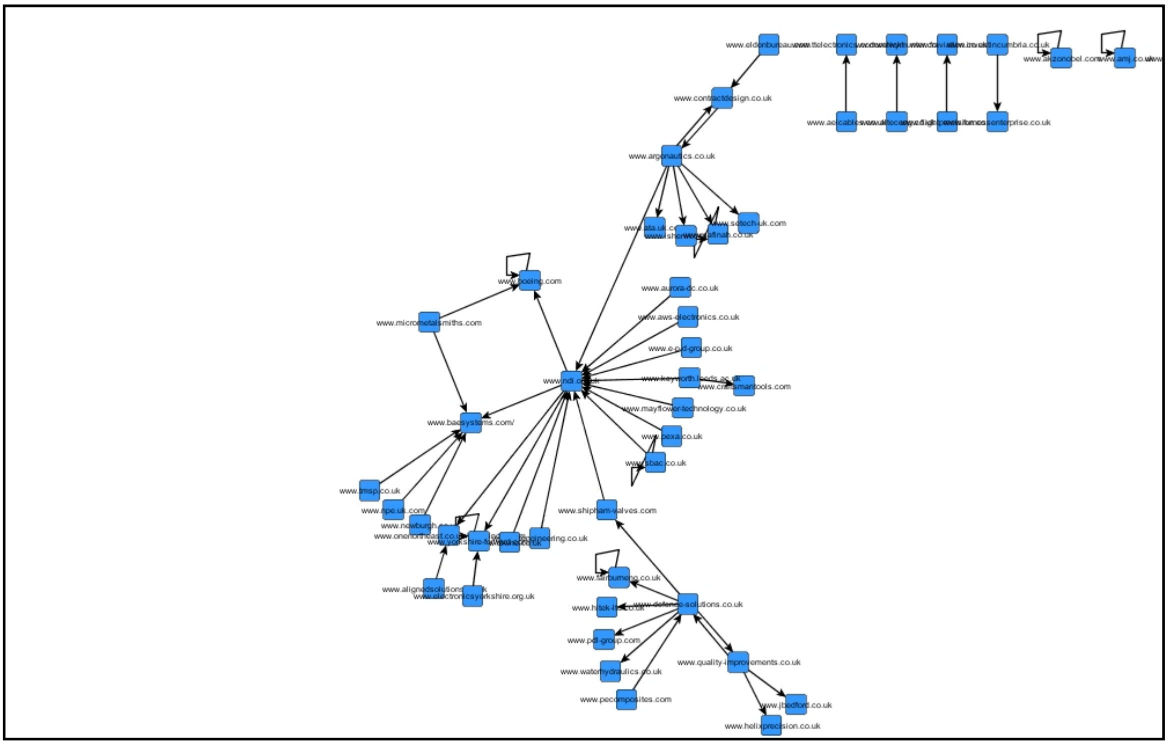


Figure 16 - Defence (detail)



Offshore - Figure 17 and Figure 18

In the Offshore segment there were 82 links out of 340 firms of which 42 were internal leaving 40 linking to external sites within the group. This is a fairly modest result as far as linking goes but the few results are interesting. Seven firms link to the Health and Safety Executive (www.hse.gov.uk) perhaps an indication of the safety culture within the offshore industry. The Northern Offshore Federation (www.nof.co.uk) itself shows up as expected but only with the firms on its publicly accessible part of the site. The Federation does have data on all its member firms but the spider would not pick these up as a user name and password is required to access this part of the site. The Argonautics group shows up again with a number of links to its members, there being a clear commonality between the skills required for defence related ship type structures and the offshore segment.

Figure 17 - Offshore (all)

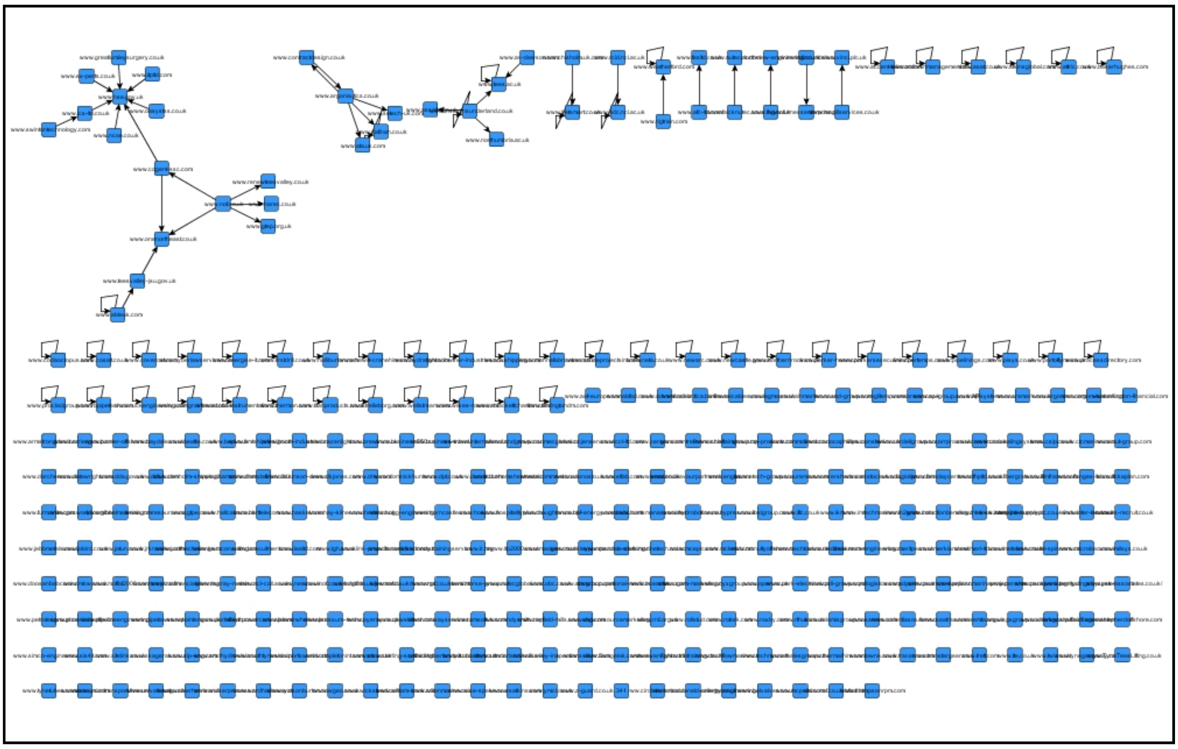
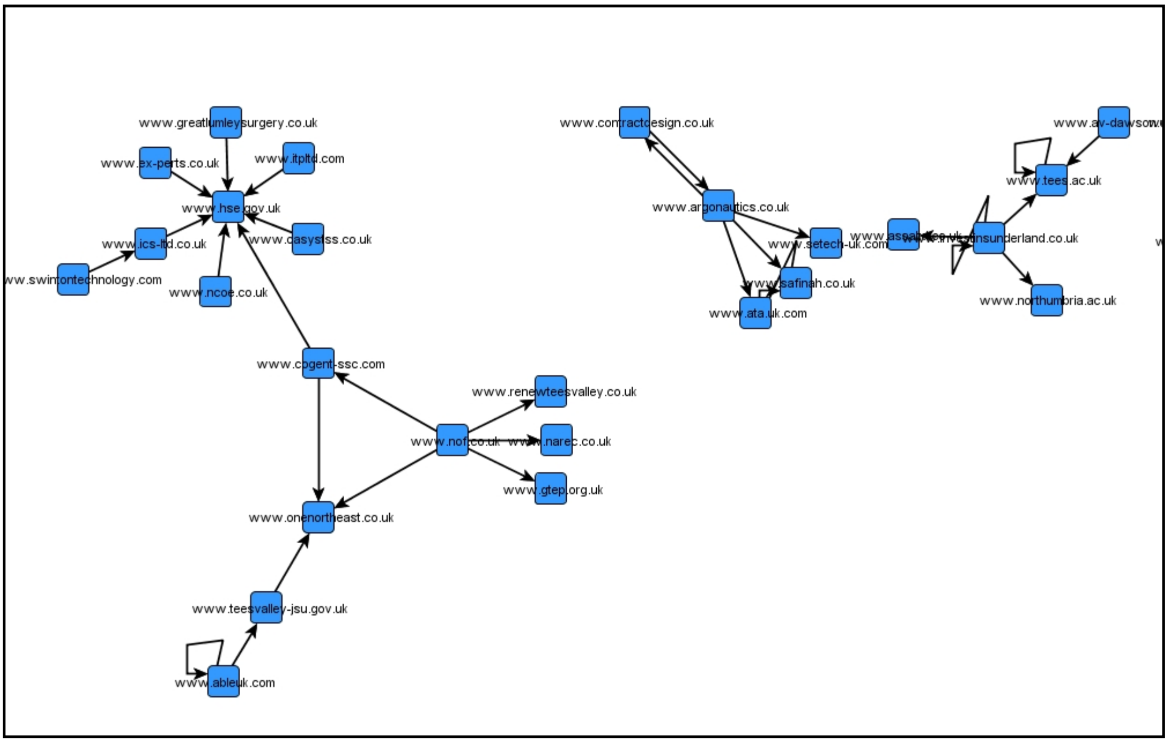


Figure 18 - Offshore (detail)



Marine - Figure 19 and Figure 20

It would be assumed that the segments of defence, offshore and marine would be similar but in fact of the 199 nodes in the marine segment only 45 had links to any other and of these 35 were circular leaving just 10 referencing other sites. There are many reasons for such apparent anomalies and some of these are discussed in the Conclusions to this section. However the principal one is that we only get results from the sample of firms we are working with and if many marine related firms have not been put into the original URL database then they will not be picked up by either the keyword spider or the link finding spider.

Figure 19 - Marine (all)

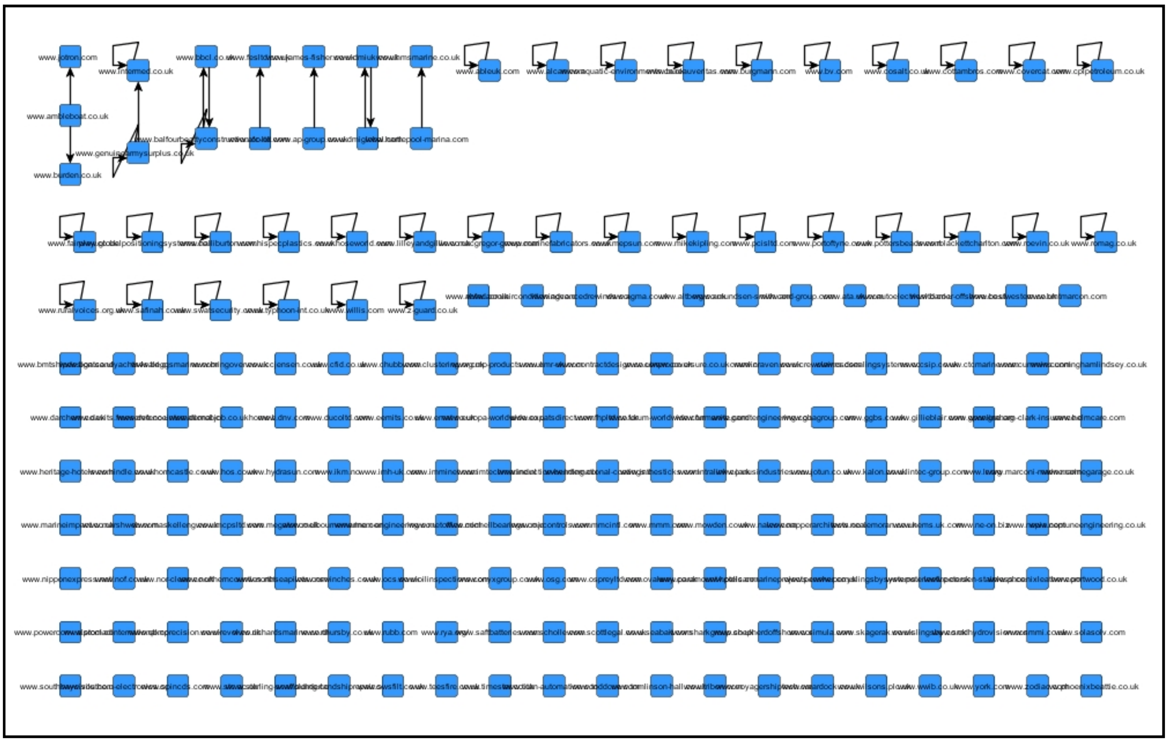
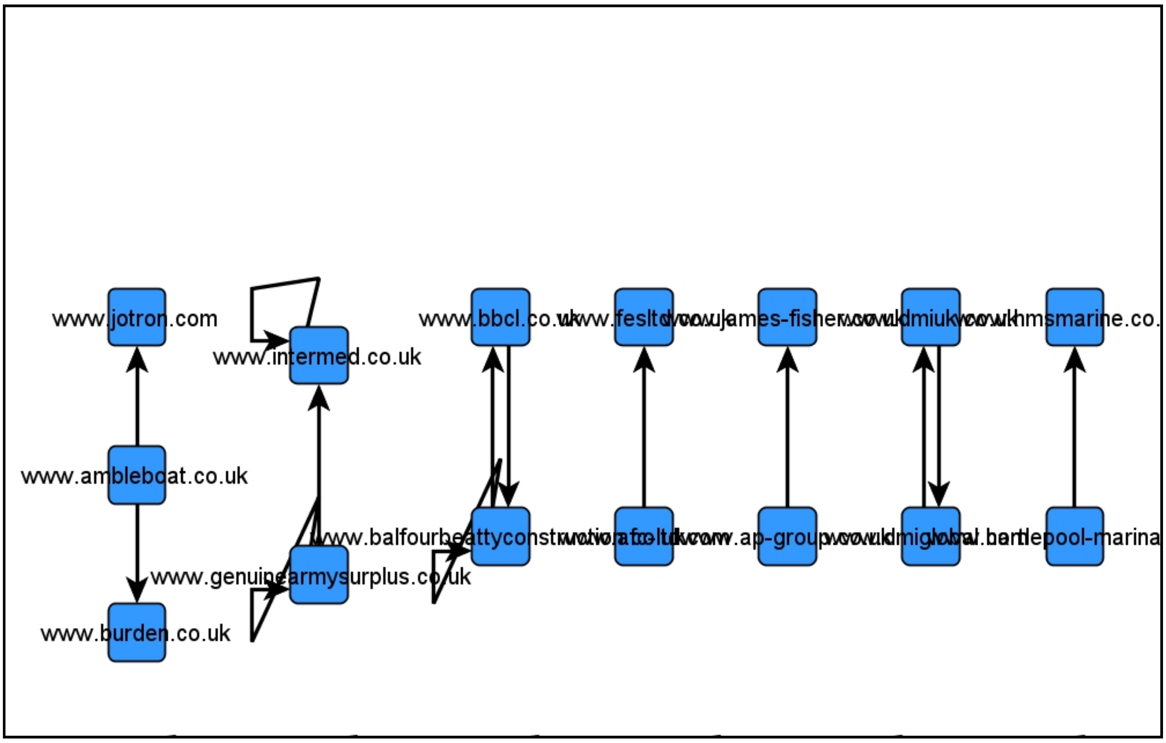


Figure 20 - Marine (detail)



Public Relations

PR was looked at on the basis that it was a more modern phenomenon than some of the segments such as 'Marine' many of whose member origins date from the last century. As such, PR firms as part of the 'new economy' associated with the internet might be expected to promote themselves and their clients might be more inclined to use their websites to promulgate linkages to their client base.

The segment not surprisingly was a small proportion of the RBKS database and numbered just 77 firms. The results were again disappointing in that there was little evidence of overt site linking to others in the same cohort and no key nodes of interest. As with the other segments examined this may be due more to bias in the original database rather than a desire to promote just themselves without reference to others. Another reason for the apparent low level of connection is evident when exploring the sites manually and that is the high prevalence of Flash driven sites. PR firms it seems, are forever trying to raise their individual profile by 'doing things differently' and as a result often resort to the extensive use of Flash on their sites. The result of this is that there are many picture driven pages but because of this the spiders are unable to track embedded links or indeed anything as often there does not appear to the spider to be anything of interest on the site.

Further Work

In some respects the results are counter intuitive in that 'old economy' groups such as Defence and Offshore seem to exhibit a greater propensity to implement website links compared to some of the newer segments such as electronics and PR. However it may be that when looking for linkages the Trade Association or focussed grouping was more effective than more generalist groups even though the latter may be segmented under a banner that suggest some similarity.

To investigate this phenomenon further the RBKS was used to look for groups with a 'new economy' bias and then others that might exhibit a very focussed or small cluster. Table 7 below shows the numbers of firms found that could be regarded as being in one of these two categories.

Table 7 - Firm count by search word

Search Word/s	**Count**
E-Commerce	146
Web	2302
Internet	1207
Hosting	495
Domain	631
Computer	513
Software	326
Knowledge	326
Consultancy	589
Engineering AND Consultancy	110
Knowledge AND services	190

Most of these searches such as for ‘Internet’ and ‘Web’ are too generic and would admit a wide range of participants most of whom are unlikely to have much in the way of linkages. However there is much interest in the N.E. region regarding ‘knowledge intensive business services’ (KIBS) hence the search on ‘knowledge’ AND ‘services’. The 190 companies found were further enhanced by the membership list of Service Network, a regional support group set up to represent and to provide a network and support for some 246 mostly small firms. When the two sets were combined and duplicates removed this cohort numbered 422.

It could be argued that in cluster research KIBS are merely the supporting infrastructure for other clusters even though Service Network as an organisation routinely refer to themselves as a ‘Service Cluster’. Nevertheless with such a sizable cohort it was felt that investigation into the degree of linking of this type of firm would be of interest. Many of the member firms are indeed the kind of organisation that did not exist as a type 10 or even 5 years ago undertaking myriad forms of consultancy, visual design and web related activity as well as more conventional services such as accountancy and advertising. These 422 were thus used as the basis for another search for linkages between them.

The KIBS segment at 422 firms is quite large compared to previous spidering exercises and consequently it took 30.11 hours with a further 0.97 hours to process the output. The results however are interesting with the overall diagram of connections shown in Figure 21 and details of the most authoritative connections (at least as far as referenced regional sites are concerned) shown in Figure 22.

Figure 21 - Knowledge Intensive Business Services (all)

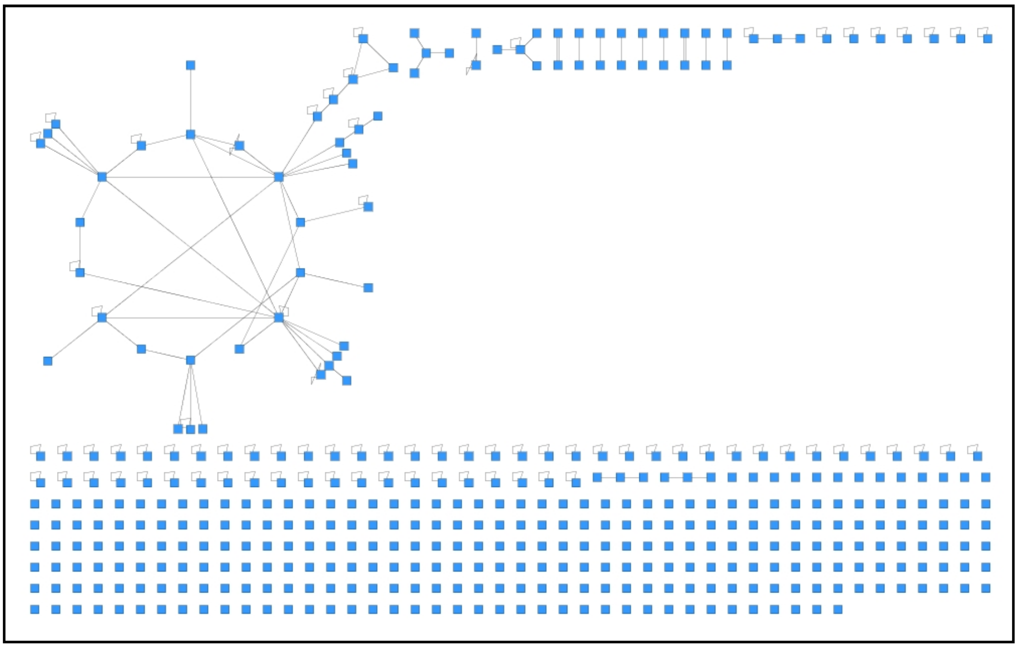
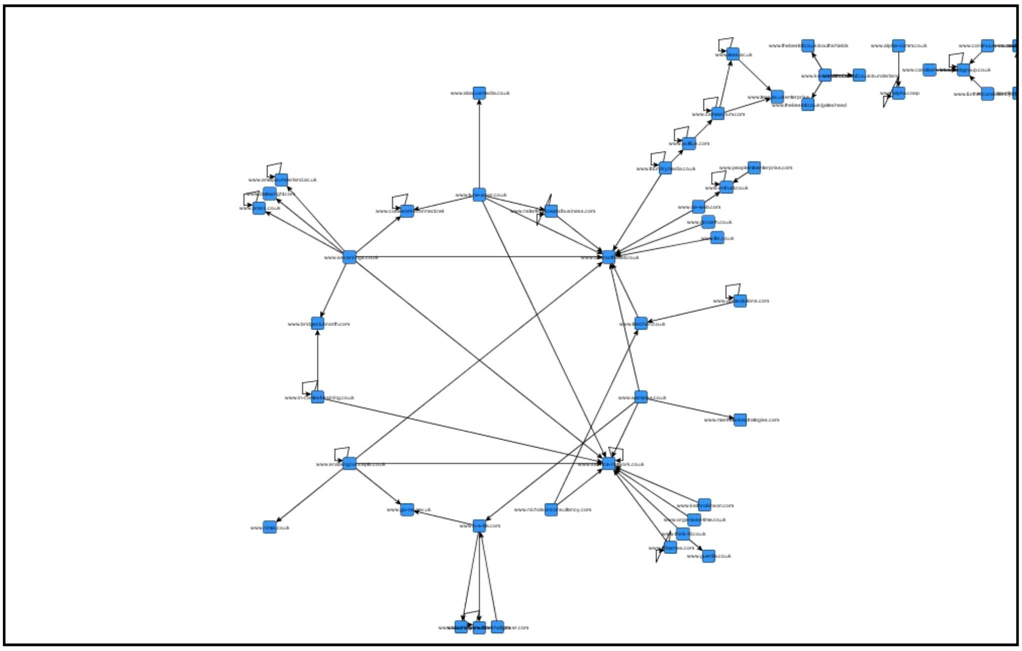


Figure 22 - Knowledge Intensive Business Services (detail)



Of the 422 nodes there were some 153 links of which 65 were internal within the firm leaving 88 linking to other sites although of these 15 also referenced their own site.

In the detail 'important' or 'authoritative' sites were www.onenortheast.co.uk, the regional development agency and www.service-network.co.uk the network membership organisation with 10 inbound links each. Only 3 firms however linked to both. Another node of interest was www.sevenrings.co.uk with 7 links. When this site was accessed it was clear that this company had in fact many more links on its site under but most of these were to sites not on the list of 422 apart from the 7 noted. In fact this company could be regarded as the complete opposite of those firms who choose not to place any links on their site. The firm has organised its site with information on other firms and organisations of interest categorised under main headings of 'Other Clients' and Bookstore and Links'. Within these categories is further information or a link to companies, public sector organisations, Education and Others with further sub-categories for Funding and grants, other consultants, Information and Networking.

The company must presumably feel that case studies with the names and sometimes clickable links of their clients is a worthwhile form of promotion but from the point of view of researching networks the URL is a small goldmine of both links and information to other clients and supporting organisations within this firm's network. A count of these links show the names of 51 clients of whom 4 have clickable links and in addition there are 36 other sites of interest to the company categorised above and all of these are hyperlinked. Thus the firm has 87 'connections' with 36 of these hyperlinked to the website of interest. The only reason only 7 showed on the spidering exercise was because of the bounded nature of the application in that the spider was only looking for connections from or to sites on the list of 422 firms and other organisations. With www.sevenrings.co.uk it should be noted that all the referenced connections are to firms or other organisations within the North East region. If of course it was possible to access a complete list of URLs in the region then all the connecting firms in this company would be picked up by the spider.

This firm is important in the context of this research because if every firm undertook to populate its website with a similar number of connections of interest it would indeed hold out the prospect of gaining a true understanding of the rich web of interaction between firms including non-trading entities and various support organisations. It would in effect start to become a reasonable alternative to the trouble and expense of site survey and interview presently necessary to discern these types of connection and discussed briefly in Other forms of Clustering on Page 38

The second category that of a highly focussed group, was subsea engineering. This has already been mentioned variously in Chapter 9 and is also referenced in Appendix 8. Chronology of URL database build. The results from a RBKS search were augmented with data from the work done by Andriani and Siedlock (2006) resulting in a cohort of 111 firms. If as suggested above from previous work, a small focussed group is likely to exhibit linkages on their websites to others of interest in the same group this should show up in the subsea cohort.

When the 111 URLs were spidered for embedded hyperlinks however the results were meagre in that only 17 links showed and 16 of these were internal to their own sites. In other words in what was a very specialised grouping no firm thought it worthwhile on their website to hyperlink to any other in that group. However further work would indicate that such a hasty conclusion is misplaced and the reasons for such are discussed later in Section 11.4. By way of top level inspection of the various firm's markets it would seem that those in the subsea market are very much globalised and an exercise looking for bounded connections just within the North East of England does lead to few candidate connections. When this is understood it may also help to explain the relatively low numbers of 'within region' hyperlinked connections for Offshore and for the Marine cohorts also.

11.1.2 Preliminary Conclusions on initial Link finding

In drawing any conclusions from link spidering at this stage it is important to bear in mind two points in the context of the research question:

i. The database of URLs is a sample and only a sample of industry, commerce and supporting activity in the region and whose sectoral bias is not fully known.

ii. Links that are discerned as a result of the scans shown in this section may owe as much to the construction and content of the individual websites as to any actual business relationships with other sites in the region.

In addition it is clear that the degree of linking varies from firm to firm and a short survey was carried out to find out why this should be so. Details of the survey are shown in Appendix 12. Survey of Firms Web linking policies. This survey whilst giving some indication of the reasons for firms not putting hyperlinks on their sites does nothing to improve the situation and it remains that most firms do not think it worthwhile to embed hyperlinks to other firms or supporting organisations on their websites.

For some of the supposed communities of practice, an example being firms or other organisations who had words or phrases containing ‘architect’ on their web site, the lack of any overt links to any other firm is counter intuitive and possibly due to cultural insularity i.e. do architectural practices see themselves as independent and competent to carry out all projects without reference to others? Similarly for firms with the word ‘subsea’ or ‘marine’ on their website again few hyperlinks exist within their respective segments. It is possibly understandable with Marine as most of that industry has roots going back over many years and the British Shipbuilding industry in particular tended to follow a combative form of subcontracting rather than the more modern concept of collaborative working. If this culture was translated in later years into individual company websites it is perhaps understandable why there appear to be few collaborative links. Conversely however the Offshore and Defence segments which have many firms in common with those also in the Marine segment do show connections and it may be that, as noted previously, networking activities of their respective trade associations have a beneficial effect on the firm’s tendency to collaborate which translates into hyperlinks placed on some company websites. It is tempting therefore to conclude that the present state of development of most companies websites contain so few links that they have little to contribute to research on networks of co-operation amongst the firm, its buyers and suppliers let alone the more difficult to track untraded interdependencies. To then make the further transition to discerning clustering activity by these methods might seem more forlorn hope than practical reality. There are however a few encouraging signs. By examination it is clear that some firms, although not having many hyperlinks do in fact give clues to organisations and the milieu of interest to the firm Such clues are in the form of text based client lists or references, logos or others that they work with.

The next section looks at some of these other ways in which the internet can be used to discern connections between firms and their industrial and service based environment.

11.1.3 Other ways of finding Links.

In earlier work and in particular Section 10.3.4, the concept of using a wide variety of means to acquire web based information on the connections that a firm makes was looked at in the context of what was publicly available. This section explores some of these to test out whether in fact they add anything to our knowledge of the connections firms make with other firms and other relevant organisations. These new tools would be an adjunct to the links found directly from a firm’s website as ‘outgoing’ links. In previous testing, the segments that appeared to have little in the way of connections were Architects, Public Relations and Subsea. As noted some of these results were counter intuitive in that it would be expected that there would be a degree of web based interaction amongst participants in the same broad segment. An obvious explanation would be simply that just looking for outgoing links in some segments just does not yield anything of value. However the view was taken that one or more of these cases was worthy of further investigation and in particular other means of finding links. Finding bi or multilateral links by spidering takes many hours of computer time, looking for links manually similarly takes time so there is little point in engaging in a large exercise if clues to why links are or are not present can be discerned with a smaller number of companies. It was therefore decided to use the ‘Public Relations’ cohort as a test as there are a relatively modest number of companies involved with virtually no web based outgoing links to any other firm in the cohort. Thus it could be argued any improvement by any method of discerning links (in addition to the few outgoing ones) would be readily apparent.

Design of the test.

This was designed on the basis of taking the 77 URLs obtained from the RBKS and which had subsequently been spidered for common regional links. The intention was to look at each of these starting at the first in the list and manually looking for inbound links using ‘Link Popularity’. This program looks for inbound links from three separate authorities although there is a high degree of commonality in the results. The search authorities are Google, MSN and Yahoo.

In addition an Alexa search for ‘Sites Linking In’ was carried out, the intention being to go through each URL and glean as many in-links as possible to each URL to add to those from the search engines noted above. Having acquired a number of additional URLs linking into any base URL the spidering program can be run again to determine if any of these add to the milieu.

Interim results.

In the event, the process of manually finding in-links was very time consuming and required a lot of attention on the part of the user and it soon became obvious that far from having a situation with a few links associated with any particular URL some had hundreds or even thousand of in-links. The initial design of the test outlined above and which was predicated on the basis of testing a relatively small number of URLs from the 'Public Relations' dataset of URLs turned out to be well founded. The sheer number of in-links found for some sites together with the usual internet related problem of lack of quality control of any kind justified the 'test it and see' approach rather than becoming involved in a major exercise to find every possible in-link for say the 'Offshore' dataset.

As a start a subset of the 77 URLs being just 16 in number were chosen at random from the PR set and all of these were examined using 4 search engines for in-links. Two URLs were removed as one was clearly non regional and the other was a dead link.

The results of this exercise are shown below in Table 8. Table 8 - No. of In-links for PR firms by search engine

	In-link source			
URL	Google	MSN	Yahoo	Alexa
www.adagencyuk.com	6	8	2782	4
www.beldons.co.uk	1	9	247	4
www.bitingedge.net	7	8	411	13
www.bradleyomahoney.co.uk	2	21	14	3
www.coolbluepr.com	0	2	2	0
www.da-group.co.uk	1340	1720	775847	26
www.enablingconcepts.co.uk	2	11	38	1
www.e-penna.com	72	999	770	40
www.fawthropmclanders.com	2	15	134	1
www.gentle-persuasion.com	0	2	0	1
www.gravity-consulting.com	1	12	16	1
www.kimmerston.co.uk	1	2	37	1
www.kinghorn-davies.co.uk	2	7	27	2
www.lexica-communications.com	0	7	7	1

As can be seen there is considerable variation in the number of in-links, as much as 5 orders of magnitude between the most and least 'popular' sites at least as far as search engine rankings are concerned. In addition there is also a similar range of disparity between search engine sources.

The former is a function of the type of site and hence its popularity. The example www.da-group.co.uk is a strategic marketing and new media company. One of their sites is concerned with a *"Sport Network having 10 million page impressions per month (and growing) to more than 500,000 unique users"*. With such a user base on this aspect alone it is not surprising that there are many external links to the sport network and hence the top level domain. In addition the company also has an e-marketing division and is presumably well aware of the beneficial effect that many in-links have to popularity statistics and hence search engine rankings.

By contrast some of the other sites may be little known, appearing by inspection to be small and local with commensurately few in-links.

The second issue, that of wide variability between different search engines discerning in-links for the same URL, is a function of how the in-links are found the process for which was discussed in Section 10.3.4. Some search engines look for links to the referenced URL within their index, others look at the 'traffic' generated and give a weighted opinion of the most popular searches to the site in question and the linking site that they come from.

In addition to the above some of the sites were looked at to see which of these in-links might be regionally based. An example shown is that for www.adagencyuk.com in Table 9 below.

Table 9 - Location of In-links

(for www.adagencyuk.com)

Inlinks (selected)	Regional	Non-reg.
http://www.attitudeswomenswear.com/		nr
http://www.bridalbug.co.uk/	r	
http://www.bsflorist.co.uk/	r	
http://www.castlegateelectrics.co.uk/	r	
http://www.croftcircuit.co.uk/		nr
http://www.croftersfoods.co.uk/	r	
http://www.douglasradburn.co.uk/		nr
http://www.grainary.co.uk/		nr
http://www.harping-on.co.uk/		nr
http://www.knickers2you.co.uk/	r	
http://www.move2day.co.uk/	r	
http://www.northernsafetyltd.co.uk/main.php	r	
http://www.prleap.com/pr/58493/	r	
http://www.stylism.co.uk/	n/a	n/a
http://www.theblackwellox.co.uk/	r	
Total	9	5
	60%	33%

The sites shown are a considered view of companies of interest to the originating company as references to catalogue and directory sites such as Yellow pages, other internal references and a vast number of Yahoo Site Explorer in-links that appeared to add little to our understanding of which organisations were 'important' to the company have been removed. Whilst such a result was entirely to be expected, after all it is the function of directory sites to reference companies of interest, in the context of this research such a feature is singularly unhelpful in the same way that such sites can be problematical when spidering sites for keywords to populate the RBKS. To remove these of course relies upon human intervention. It would be possible not to have human intervention at this point in the same way that Zoominfo© (www.zoominfo.com) outlined on page 164 looks for existing associations when given a company or person's name and automatically links them to the subject. As noted however Zoominfo does ultimately rely on an individual to correct any inaccuracies in the automated collection process for their entry. At this stage of development of internet tools it would appear that human intervention is a necessary part of avoiding the kind of largely irrelevant sites just described above. Similarly to determine whether a site was regional or non- regional ('r' or 'nr) in

Table 9 above for a relatively small number it was a matter of visiting the site and looking in the ‘Contacts’ or ‘Contact us’ section which is present on almost all websites. For a very large cohort the option exists to resolve the IP address geographically and compare this with a reference area but again the presence of catalogue sites with an IP address derived location far removed from the actual firm location makes this whole process a less than perfect exercise.

This subject site (www.adagencyuk.com) was a company located in the south of the N.E. region with what appeared to be mostly local clients. Although 60% are shown as regional of the 33% shown as non-regional in fact most of these were located a short distance away but outside the regional boundary. For completeness the one site (7%) shown as N/A was merely a third party inactive holding site for that URL.

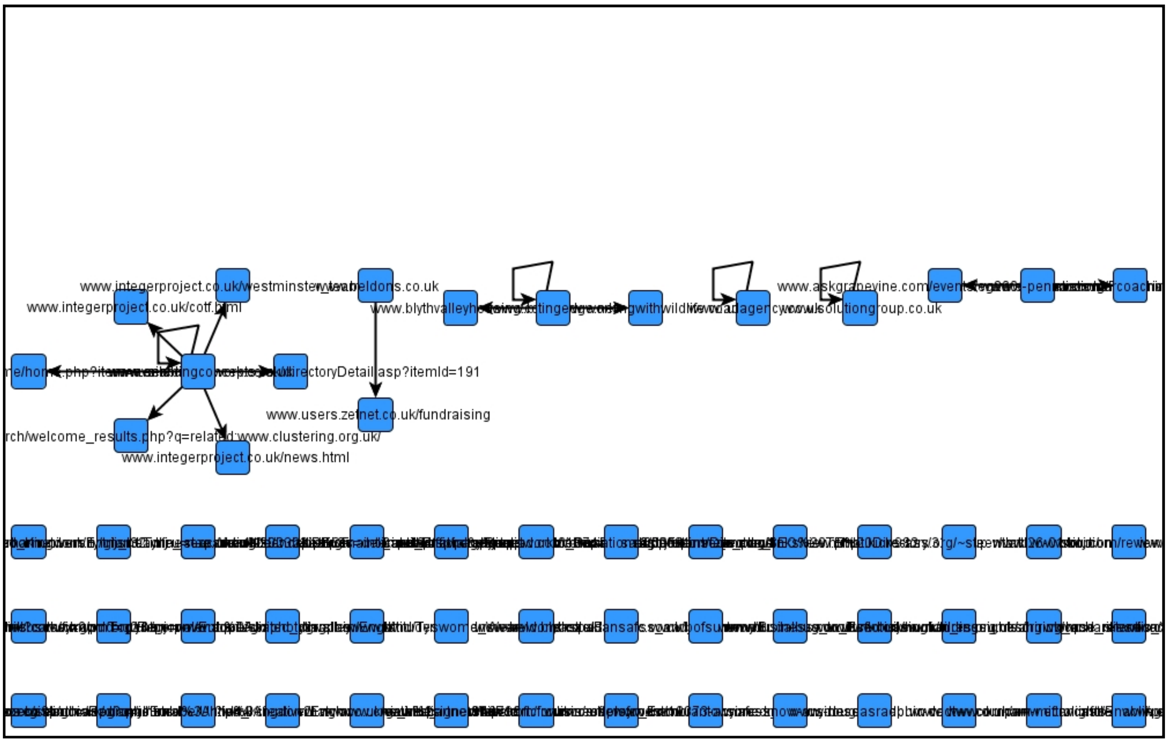
After the 14 sites noted above had been examined it was clear that the time required to examine even a modest cohort by searching all 3 link popularity programs plus Alexa was significant. As a result of this it was decided to complete the Public Relations segment just by using Yahoo Site Explorer and taking a view on the results obtained for each URL by removing the unsuitable or unhelpful linking sites previously noted. Whilst this cut down the time required the exercise was still non-trivial for time invested as a function of the number of results obtained. Generally, as with all search engines, on the basis that early results are the most relevant and as each page returned up to 50 in-links, only the first page was used. Some URLs of course had few in-links but again only ‘relevant’ ones by inspection were included. Yahoo Site Explorer does have an advantage in that its results can be exported into a TSV file which can then be read and manipulated in MS Excel. It is also possible to look up the referenced site direct from Excel to check the validity and physical location of the subject site.

It is worth noting that during this process it again was demonstrated why PR firms appeared to have a low number of connections and this was due to the large proportion of ‘Flash’ driven sites. This reason has been alluded to before in general terms but it would seem from the small sample here that at an unusually high proportion (compared to the general URL population) operated sites that had at least a Flash front end and often much of the content as well. Such sites make it difficult for search engines to index anything and although the sites may have embedded metatags (which would be picked up by the RBKS spider which identified the site with the words ‘public relations’ in the first place) there is still the problem of reduced visibility. Theoretically this should not affect externally generated in-links as third party sites can still reference flash based sites.

Finally for Table 8 all the URLs obtained as part of the in-link acquisition process plus the 14 originating sites (184 URLs in total) were spidered for associations. It would of course be expected that at the very least the originating URL plus an associated in-link would be found. It should also be remembered that when the original 14 PR subset was spidered NO links at all were found so it could be argued that any links found would be an improvement.

The result of the exercise is shown in Figure 23.

Figure 23 - 14 PR firms with in-links (detail)


www.integerproject.co.uk/westminster_
www.integerproject.co.uk/cotf.html
www.sheldons.co.uk
www.blythvalleyh
directoryDetail.asp?itemId=191
www.users.zetnet.co.uk/fundraising
rch/welcome_results.php?q=related:www.clustering.org.uk/
www.integerproject.co.uk/news.html
www.askgrapevine.com/events
solutiongroup.co.uk

Interim Conclusions on the use of in-links.

The objective of this section was to determine if acquiring in-links would be possible and if the results obtained would be of any benefit in the overall context of finding links between companies. There are three observations from the findings above;

1. In a perfect world it would be expected that for each of the originating URLs, when these were spidered in conjunction with their found in-links there would be as many associations from the spider program as there are in-links in the dataset. This is not the case as the number of associations found are less than the number of links determined by the in-link finding program Yahoo Site Explorer. Possible reasons for this phenomenon are considered later.
2. Many of the in-links in the dataset may be from outside the region.
3. In spite of the above shortcomings there are clearly more links and hence associations between companies than mere searching of company websites.

It could be argued therefore that this exercise is a limited success in that it shows the use of programs to discern in-links helps to find linkages between companies that were not obvious from overt hyperlinks on a company's website.

In particular Figure 23 shows a single company emerging as an 'authority' in the manner of Chakrabarti et al (1999) and is a similar to those key nodes shown in Figure 22 which were derived from overt site hyperlinks based wholly in the region.

One of the drawbacks with using in-links to augment the firm's associations is that at the present time this has been carried out manually requiring significant effort to both undertake the task of capturing these in-links and then making a judgement on what is an 'appropriate' in-link to use. The time is further extended if it is required to determine if the in-link is from a regional company as this is usually done by examining the associated website. As noted it may be possible to automate these processes to some degree by using a filter to take out any catalogue, directory or other nominated type of inappropriate site. The location of the in-link site could be dealt with by resolving the geographical location of the IP address. There are significant difficulties in making this work for all types of in-link in the same way that the regional spidering program, after many iterations and developments is still occasionally defeated by some unusual website internal construction as commented on page 187 onwards. An example might be a reference to a company in a newspaper article which would be picked up by Yahoo Site Explorer searching for in-links. Clearly the news article would not be a linked company and the location of the news article derived from the IP address for the news agency would be irrelevant.

11.1.4 Links by Inspection

Thus far examples have been given, in a cohort of firms, of hyperlinks discerned by a spidering process (outgoing links) and those found references from other third party sites (in-links). There is however a very simple method previously alluded to that can find other outgoing linkages and this is by simply looking at the firm's website. Whilst this process may seem rather pedestrian compared to the more esoteric processes of spidering or using the very complex science behind some of the in-link finding programmes anything that contributes to answering the research question is to be valued. With this in mind a further test was set up to look at the contribution to the finding of associations that could be made by visual inspection.

As noted above, in previous testing the segments that appeared to have little in the way of connections were Architects, Public Relations and Subsea. We have already examined Public Relations in the context of in-links and noted that the number of connections found is improved. Whilst it would be logical to go ahead with finding other associations on the same cohort it was decided for this test to use the Subsea group. The main reason is that a great deal of work had been undertaken by Andriani and Siedlock (2006) on the emergence of subsea activity in the region and indeed their paper was entitled 'The Emergence of the Subsea Cluster in the North East'. The authors at least felt that there was a cluster in subsea activities and the attraction here is that the cohort, although small in Porterian terms, would serve as a useful comparator.

A further advantage to using the subsea cohort as a comparator was that, in correspondence with the authors of this work, a diagram was obtained of the interconnections between firms that they had found by snowball sampling and interview. These interconnections were derived on a dual basis of trading relationships and/or of knowledge acquisition and it was felt that the 'shape' and density of the networks seen would be useful to compare with a similar network derived by internet based means alone. In other words if by using internet based methods, the numbers of firms found, their capabilities and their interconnections looked anything like a subsea network as already discerned by these authors then that would be useful knowledge and would indicate, for the subsea cohort at least that it was possible to obtain associations between firms and other organisations without having to undertake face to face surveys.

Design of the test to determine additional firm links and associations

Thus far in the research much has been learned about the capacity of the internet to provide useful knowledge about firms and their activities. Conversely the many shortcomings in such an approach has in many cases been readily apparent. This part of the investigation is slightly different in that we are proposing to manually examine a cohort of 111 websites identified by means already described as being part of a subsea cohort. By examining the sites we will be looking for any reference to any other site that may be in the region.

The subsea cohort of 111 was established from two principal sources being the Siedlock and Andriani (2002) database and the RBKS. In the original data made available by the former some 103 firms were identified. This identification did not however include URLs and these had to be determined manually using a search engine. This resulted in only 80 firms being identified with a valid URL the remainder appearing to be small subcontractors without a website at the time of searching. The RBKS was also used to find firms with the word ‘subsea’ on their website and after removing duplicates already appearing on the larger list but adding few with internal embedded links the working list of 111 was obtained.

A few ground rules were established to guide the subsequent process as the firms and other organisations varied in size from a small local firm to very large multinationals but who had a local presence such as a branch office, plant or construction site. The basic precepts used were:

- Sites to be scanned for overt references to another company or organisation such as a university
- Scans to made of client lists, supplier contacts, trade associations and partners
- Scans to include supporting institutions, local organisations that helped or came into being as result of the cohort
- A ‘local’ presence was required in found links. This would therefore include any firm domiciled outside the region as long as it had a local office or manufacturing plant. It would also include some long term projects located in the region e.g. Shell oil platform.

- News publications were generally omitted but not relevant specialist publications if they had a local presence or exceptionally a website that was clearly acknowledged to be of use to local industry e.g. subsea.org
- Anything else that would add to the general desire to find connections between and amongst firms and other organisations to be included.

The method of approach was to firstly to manually access a single URL on the subsea list and to page through using the guidelines as above. For each site any external company reference was noted. As this was usually in the form of a name or logo, Google was used to trace the URL which was then looked up to find if it was regional company. If it was seen to have a regional address (even for a branch office) the company name and/or URL was added to a sub-list associated with the URL being queried.

In addition to the obvious company linkages such as client and supplier lists there were others of interest not immediately obvious. For example many manufacturing and design firms in the offshore business (and here most of the firms in the subsea cohort were part of a wider offshore grouping) have some sort of formal quality accreditation. Because of the scale of activity involved and the need to have subsea vehicles and equipment constructed in compliance with Classification Society rules, many of these Societies have an office in the region. The example seen most often was that of Lloyds Register Quality Assurance (LRQA).

A further source of company connection, although this did not occur too frequently was that of Venture Capital and of the personal links of Directors. The CVs of Directors (particularly non-execs.) on websites had names of other companies that they were associated with, often in the same sector.

Observations and Results

The original 111 firms reduced to 108 as 3 sites were not functioning. Of these some 39 were able to yield useful regional links of one form or another. Of these 39 the lowest number of links was 1 i.e. a single reference to another organisation whilst the highest number was 17. The average number of links for the 39 firms was 5.5. In total there were now 213 links which was almost twice the number on the original list. However whilst this was very encouraging it was observed that potentially many more links could have been available. The reason is that very large organisations such as AMEC (www.amec.com) have a substantial website covering the company's operations in scores of countries. They do have a client list but it refers to large organisations for whom they have undertaken a multitude of projects but often even the country is not specified let alone the North East of England. In the offshore sector (and other construction related sectors) AMEC are a major player and have worked with and for all the other majors in oil platforms, subsea construction and supporting infrastructure. Consequently when traversing the site there is very little that can be discerned when trying to gain knowledge on their local connections. This a major disadvantage which is discussed later as often subcontracting networks form around a small number of significant main contractors and if these large players are all but invisible to an internet based search for links then unless there is some other way of discerning their presence their key place in the network may not be seen.

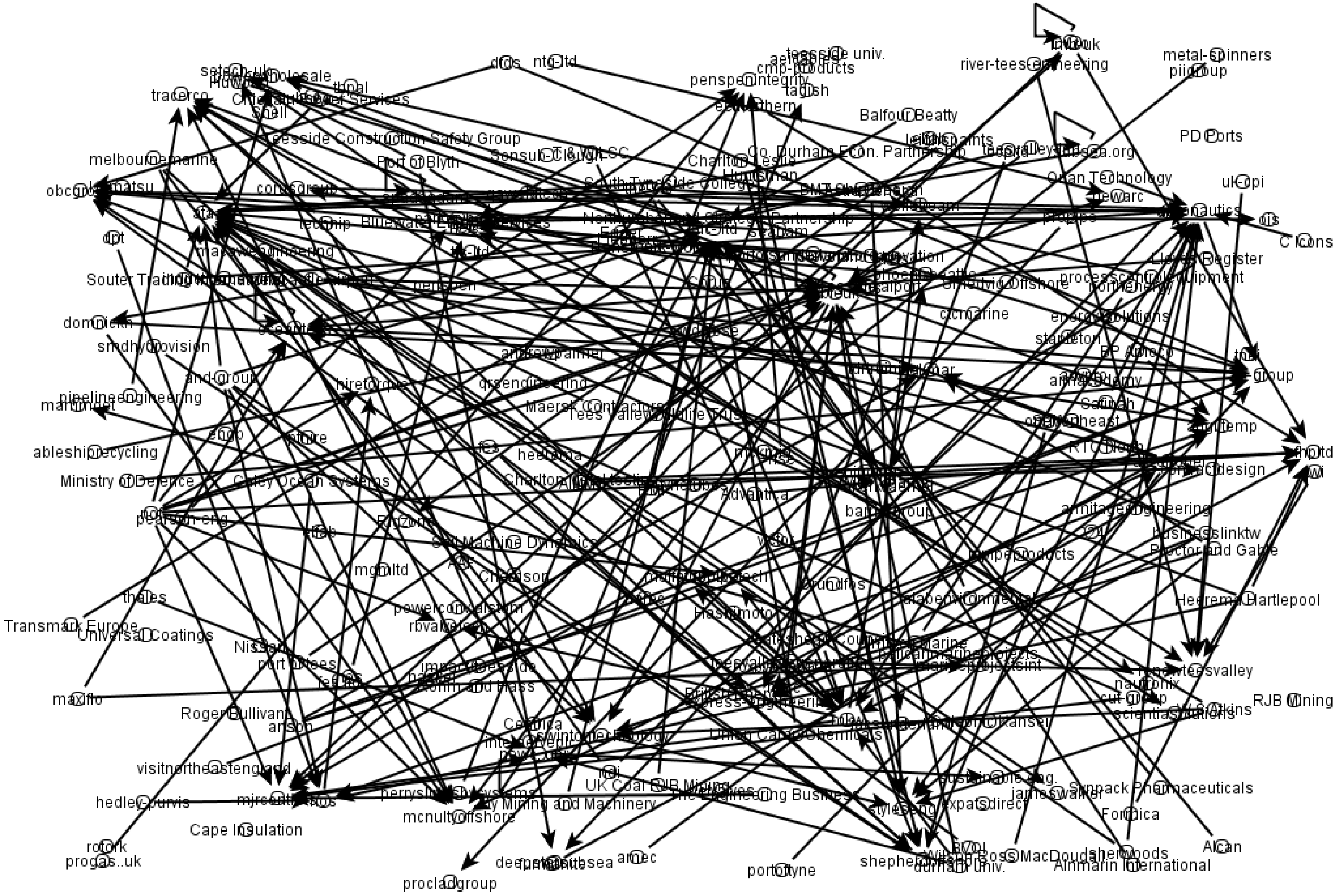
This leads to a second class of problem with the subsea cohort and also the offshore segment in general in that company websites often refer to their client by project rather than by originating client. An example would be the 'Bonga' Floating Production, Storage and Offloading (FPSO) facility converted on the River Tyne using AMEC's yard. The term 'Bonga' refers to the operational station which is located in Nigeria. Although such a large project involved many subcontractors from around the region on AMEC's site, data on the conversion is limited to a press release buried many levels down in the 'Projects' section. The same applies to other large main contractors who have undertaken a number of projects on the Rivers Tyne and Tees including some of those noted above but it is simply not possible to discern the location easily from their sites.

A third issue is the global nature of the industry. With companies involved in both Offshore and Subsea, although they may have started in the region with regional markets it was clear from looking at their websites that the majority were involved with exporting both expertise and manufactured items. As a result many of the websites emphasised the global nature of their business rather than local connections. A further example of this kind of difficulty was the former Hedley Purvis (www.hedley-purvis.com) now renamed Hydratight after being bought out by the US based Actuant Corporation. What was once a regional company albeit with global markets is now just one of 27 locations worldwide providing bolt tightening technology as part of a global conglomerate in a variety of markets that also include offshore and subsea. The company's services are the same as before the takeover, it is just that their place in the corporate firmament and hence the website does not feature local links as it once did when it was an independent venture.

The additional information on inter company and other linkages gleaned via company websites using manual inspection is several orders of magnitude greater than those gained by spidering a bounded set of URLs in this particular case. As always the issue of website design and the connections within reflecting any sort of trading or knowledge exchange relationship comes into play. As has been commented on previously, the export oriented and global nature of subsea activities does tend to militate against finding local connections particularly as far as the big players are concerned.

The results of these new connections, which in some ways are analogous to reading a company brochure for evidence of association with other companies and organisations have been graphed in the usual way. The first is Figure 24 which uses a random graph and it is immediately obvious that for the subsea cohort the density of network connections is significant compared with all the others that have been examined so far, indeed when the first set of URLs (then 111) were spidered so few links were found that it was pointless drawing any sort of a network diagram. In this case however it is clear from many common connections that there is an interesting network and that popular nodes and authorities would best be displayed by examining such a diagram.

Figure 24 – Subsea: Embedded hyperlinks and observed connections (Random Network)



This network gives an indication of the density of the connections between firms but is of itself such a dense mesh that the true topography is difficult to discern and the most popular authorities can only be guessed at. However by using either a circular mesh as in Figure 25 or a uniform one as in Figure 26 the isolates (those organisations with no connections) are displaced from the main body of the mesh and the resulting figure gives a much clearer picture of which organisations are important in a network sense.

To be clear about the network as portrayed; it is derived from a combination of sources being:

- Companies whose website has the word ‘subsea’ somewhere in the text or as a metatag.
- Companies who appear on the Subsea list as determined by Siedlock and Andriani (2006) and who had a searchable URL.
- Internal URL references from one site to another within the list as derived from the two sources above. These can be regarded as out-links.
- References to other companies or organisations obtained by manual scanning of the websites of the subsea cohort provided such references are with respect to regional location for activity, domicile or representation. This group are also out-links.

It can be seen that there is significant referencing ‘activity’ but it is of the form whereby many firms have their own clients or others with whom they interrelate but they also have some connections with the wider milieu. This sort of thing is particularly apparent in Figure 25 – Subsea: Embedded hyperlinks and observed connections (Circular Network) in that the topography is such that single firms around the edge have their own links, usually to clients and some suppliers but also a link to trade associations such as nof Energy, www.subsea.org and other groups such as Argonautics (www.argonautics.co.uk). This latter group was set up in 1990 with help from the local Council in North Tyneside organisation through a since defunct organisation called the Regional Centre for Clustering (RSC). The RSC was specifically formed to broker co-operation between same sector companies grouping together to address wider markets that individually they could not contemplate. Although the RSC was promoted at the time as an vehicle for forming ‘clusters’ the members in any particular group were only ever a few in number, had a non-competition policy internally and could not have been regarded as a larger scale functioning cluster [Pellow (2005)]. Nevertheless it can be seen that with the passage of time the Argonautics group has become an authority as far as the network for Subsea is concerned with members showing prominence with links not only amongst themselves but also to the wider network

Other firms are relatively isolated, at least as far as their web links are concerned in that they seem to operate quite independently with just their own clients and suppliers. There is also some evidence that the related and supporting industries which are harder to capture in SIC based cluster studies are starting to emerge. Apart from connections to Universities it is clear that subsea firms have links with the regulatory authorities and regional organisations aimed at promoting business and industry. In addition there are the classification societies of which the previously mentioned Lloyds Register Quality Services (LRQS) would seem to be the most popular (although it could be argued that the paid for element of both Universities and Classification Societies class them as a trading entity). The various shades of the term ‘non-trading dependency’ are discussed later.

There is also evidence of other firms that at first glance would not be expected to be part of the Subsea network and these relate to some of the process engineering and mining industry firms active in the south east Northumberland area. An example of a key firm here is Ashington Fabrication Ltd (www.afc-ltd.com). The latter sector is particular interesting in that it became clear from early work on the keyword based activity that for example, many general engineering firms sell into a variety of markets such as Offshore, Defence, Process and Construction. What may be showing up here with the work on linkages is an example of firms who came from a mining machinery background, moved into Offshore and then into Subsea. An obvious spur for this would have been the precipitous decline of the coal mining industry in the region and nationally over the last 30 years but also the skills and capability of such firms to deploy their experience of say hazardous environments with a high safety culture e.g. intrinsically safe lighting into the offshore segment. This type of transition bears on the work of Boschma et al (2008) regarding ‘related variety’.

There are also some firms which appear to be a locus of activity but the reasons for this seem at best counter intuitive. One such is Able UK (www.ableuk.com). This company is well diversified in that it has a number of businesses in offshore construction, reclamation, waste disposal and property. Quite a few links on the site are to its own subsidiaries. It was noted in the 'ground rules' that news publications were to be generally omitted but not relevant specialist publications if they had a local 'presence'. The reason for this bears on Able UK's activities in that for the past 3 years they have been embroiled in an argument with local government and sundry environmental groups regarding the dismantling of a number of ex US naval support vessels at their Graythorpe yard near Middlesbrough. As a result of this there are a significant number of both outward and inward links from a variety of news sources chronicling the saga of the company's fight to be allowed to dismantle the vessels. In terms of a subsea network AbleUK is undoubtedly a contributor but taken at face value its web connections, particularly inbound, are being dominated by the press and other noise associated with environmental issues. A similar situation occurred with Melbourne Marine [www.melbournemarine.com] when they became the centre of attention regarding the environmental implications of proposed ship-to-ship oil transfer operations in the Firth of Forth. The resulting plethora of press articles, legal submissions and reports dominated the results of in-links searching and required manual sorting to filter out what was appropriate in the context of valid industry connections.

As a final note when the original nodes and connections were combined and duplicate links removed the number of nodes (organisations) had more almost doubled to 216 with 203 linkages between them.

The comment thus far has not included any substantive work on in-links derived from other external regional organisations referencing the subsea cohort on their own website and the logical step to gain an understanding of both external and inward pointing links was undertaken next.

Figure 25 – Subsea: Embedded hyperlinks and observed connections (Circular Network)

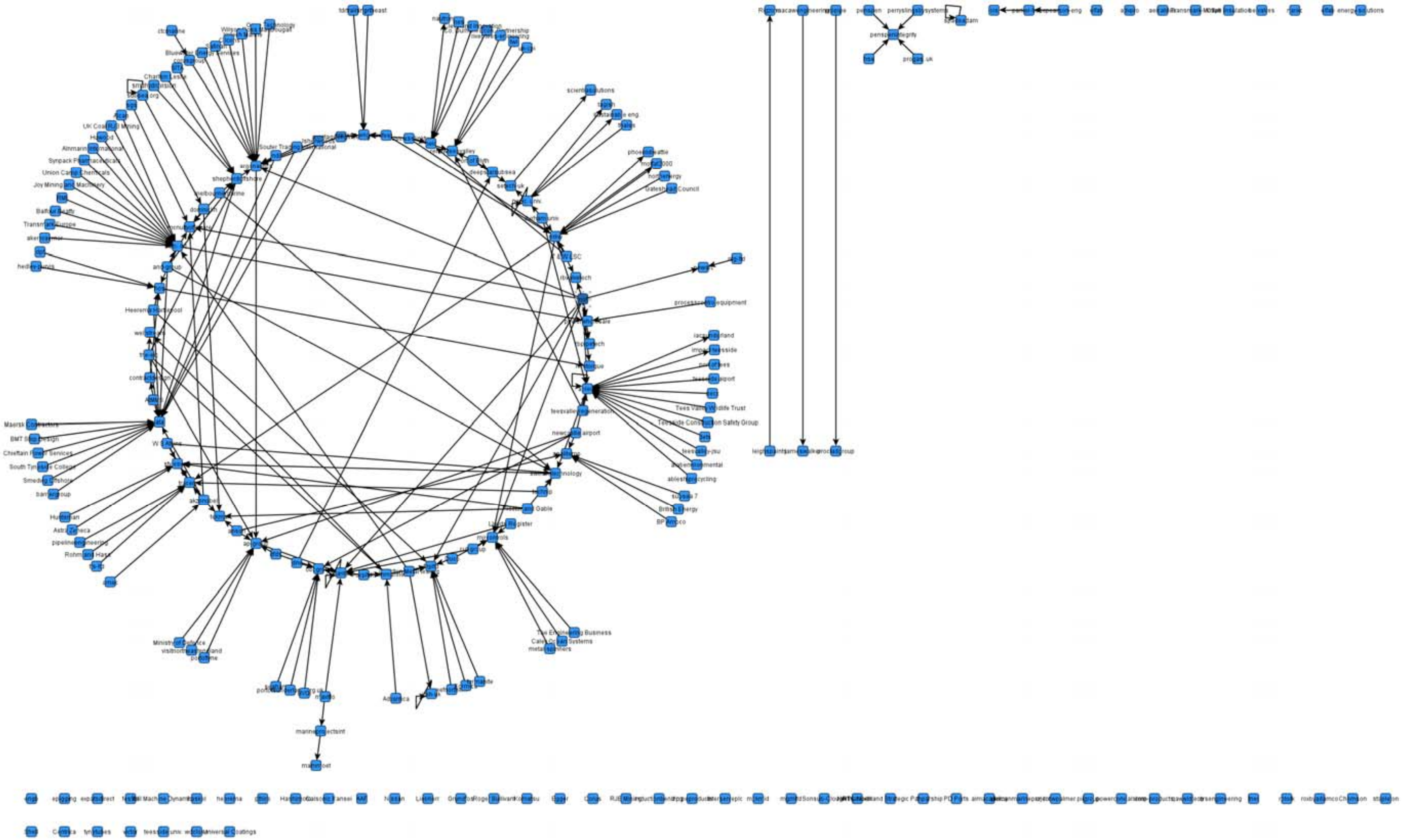
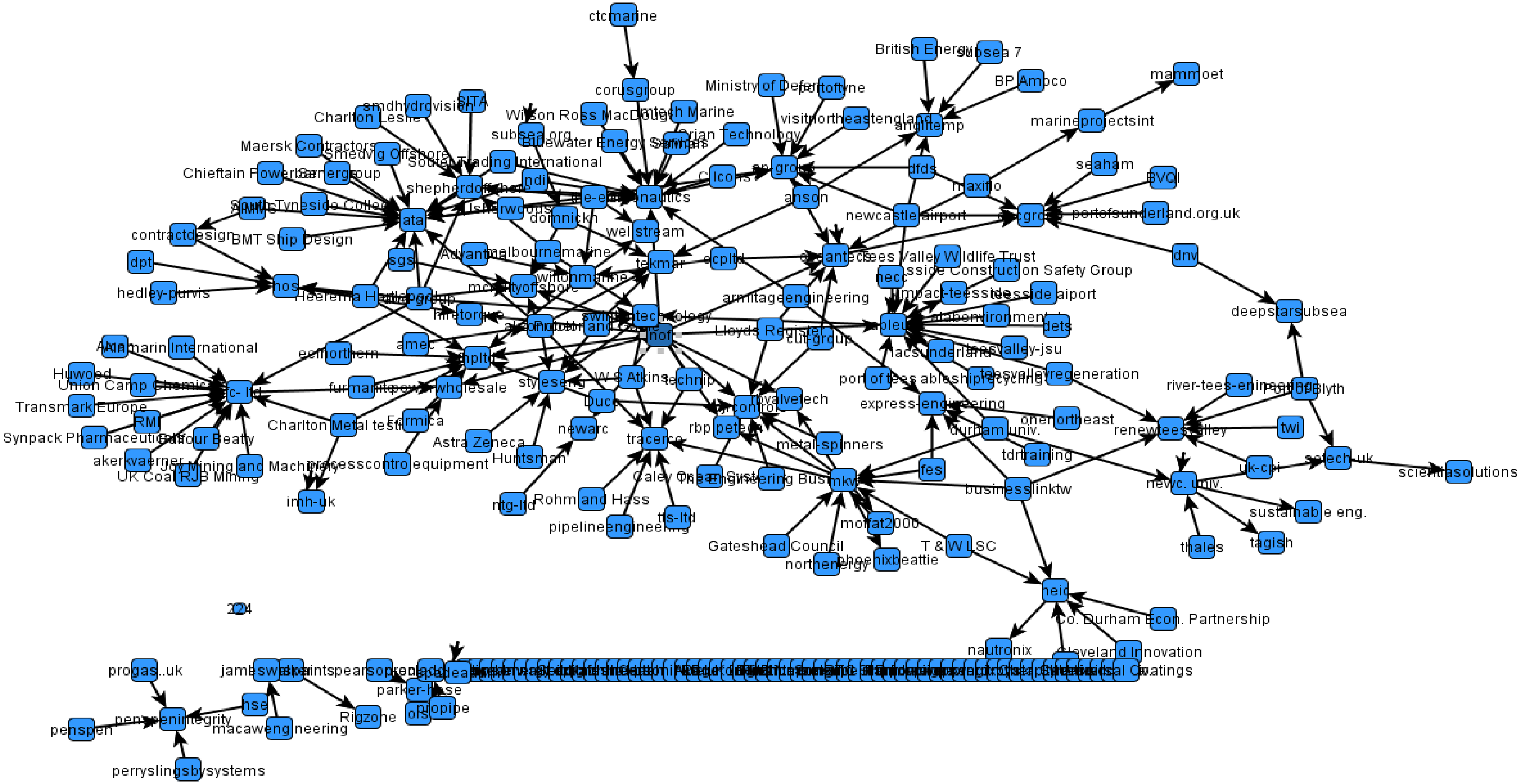


Figure 26 - Subsea: Embedded hyperlinks and observed connections (Uniform Network)



11.1.5 The Acquisition of In-links for the Subsea Cohort

The process for acquiring in-links for a small test set of URLs has been described above with some interim conclusions on the subsequent results shown in Section 11.1.3 on page 208. The process for acquiring in-links for the Subsea cohort followed the same process but with a much larger base population of companies being 238 in number which is the original 108 plus the additional organisations identified as a result of that exercise in Section 11.1.4.

It was known from the tests that had been done earlier that the use of www.linkpopularity.com was an efficient entry point to 3 independent link searching engines from Yahoo, MSN and Google. Of these, in the tests Yahoo invariably returned the most in-links to a given URL with MSN next with Google generally having the least.

For each company URL therefore Yahoo Linkfinder was used to find the first 100 in-links to the URL being accessed. Each in-link was scanned to determine if the referenced in-link was from a company or organisation with a regional ‘presence’. The term ‘presence’ has already been discussed in the context of other outgoing links in that there had to be some regional connection. This was of course a time consuming process as unless the in-links were from an organisation known to the author the in-link reference URL had to be manually opened and the website scanned usually for ‘contact’ information.

Some sites had many more than 100 in-links, others only a few. For the sites that had more than 100 (sometimes several thousand) in-links, to undertake manual inspection of all of these, apart from being inordinately time consuming would also have been less than worthwhile as there appeared to be a ranking system at work in the search engine which gave the most appropriate results first. In the event that there were only a few in-links discerned by Yahoo then MSN and Google were also searched. In practice the overwhelming majority of in-links were found by the Yahoo engine with MSN adding some extra. Very few came via Google and these results are consistent with the tests on smaller samples carried out earlier.

MS Excel 2007 was again used as the medium for capturing these results in that against each base URL all the associated in-links found for that URL were stored as an embedded and numbered sheet. An example of the top sheet used as the guide to all the other groups of in-links is shown in Appendix 9. Example of Tracking Chart for In-links. The results were then collated in a separate summary sheet constructed in .csv format as input, initially to the graphing program VISONE. It was also important to have the links 'pointing' in the correct direction as when using VISONE to show the linkages it has the facility to portray 'directed' links. In the case of the first exercise the links from internal embedded references and by those gained by inspection are all outward looking whereas with in-links the links are incoming being referenced by external sites.

Observations and Results

The result of this exercise was that the 238 nodes examined for in-links had 252 connections within the cohort.

As has been noted but which is worth emphasising, there is a fundamental difference between the embedded and external reference derived links of the previous section and the in-links which are part of this exercise. Although the patterns of connections might appear similar at first glance (apart from the direction of the link) the basis is different.

As an example it was noted that with external links many large companies did not have references to third party organisations such as clients overtly shown on their sites and the reasons for this have been discussed. With in-links however many firms who have done work for the same larger company will reference them on their site perhaps as clients or projects and this is picked up by the spider looking for references to the large company URL in question. Thus Shell http://www.shell.com/home/Framework?siteId=uk-en whose website is a typical corporate 'look at me' site with an attempt to cover its global operations in many sectors together with sections on corporate social responsibility, shareholder information and so on would never have space to reference a small local outfit that did some work for a second tier subcontractor who was in turn supplier to a Tier 1 organisation responsible for some major element of the Bonga oil platform conversion on the Tyne. That same local contractor however would be very likely to reference such a prestigious project even though its contribution may have been modest in the overall construction scheme. It is just such references that are picked up by the search engines of Yahoo, MSN and Google and which allow us to gain at least some insight into the connections within that part of industry.

There are of course difficulties. Staying with the Shell example, the Bonga project was not primarily a subsea project as it was concerned with the conversion of an existing large floating structure to the role of a floating oil processing and storage system for offshore Nigeria. However the word ‘subsea’ must have appeared on a relevant website somewhere, either Shell’s or the small contractor’s and examination of the technology deployed by the Bonga project shows that the Bonga vessel, as part of its operational role does in fact connect with a significant structure placed on the seabed. Such subtleties are difficult for the researcher not imbued with industry knowledge to discern, particularly if undertaking SIC based research but by using methods such as in-link searching it is possible to gain an understanding of the intricate web of subcontracting networks that, in addition to well known industry contributors also uncovers other firms that may not be expected to appear This is a theme that is returned to in the overall conclusions at the end of this section.

As with the graphs for the networks derived by embedded and links by inspection the three forms of graphs being random, circular and uniform but obtained using in-links are shown in Figure 27, Figure 28 and Figure 29. The names of the nodes and in some cases the direction of the arrows have been removed for clarity so that the linkages near nodes may be more easily observed. On the live network it is an easy matter to expand the detail and to click on any node to determine its attributes including the name. The program can also show the most common metrics for each node such as centrality with indices of pagerank and authority. It should be noted that there can be a number of measures regarding the term ‘centrality’. Here the nodes and links are simple with no ‘strong’ and ‘weak’ linkages and all nodes are given the same weight or importance prior to assessing their popularity. In observing links derived from the internet it is not generally possible to discern the magnitude of the esteem in which one organisation holds another but only who has been referenced on a particular website. Because of this the measure of centrality here has been taken as a simple count or ‘degree centrality’ further spilt by ‘in-degree’ or out-degree’ with respect to the direction of reference and the arrow shown on the network maps. This theme is returned to in Note on the use of Social Network Analysis and graphing programs on page 246 and in Section 11.4.

In some respects the topography of the networks is similar to those derived for the previous set in that there are a number of hubs and authorities, there are some unconnected outliers and there are firms that appear to have their own small eco system. In this case however it must be remembered that the ‘popularity’ of any node is driven by incoming links and not by links that the firm has decided to put on its own website showing a collaborator or site of interest. In some respects the derivation of the ‘importance’ of nodes with many in-links is analogous to the more sophisticated methods in use by search engines to rank websites. The examination of in-links only whist being of interest in its own right is a key part of the combined set of linkages that go to make up the total found network of subsea involved firms and other organisations and the complete picture is discussed in the next section.

Figure 27 - Subsea: In-links (Random network)

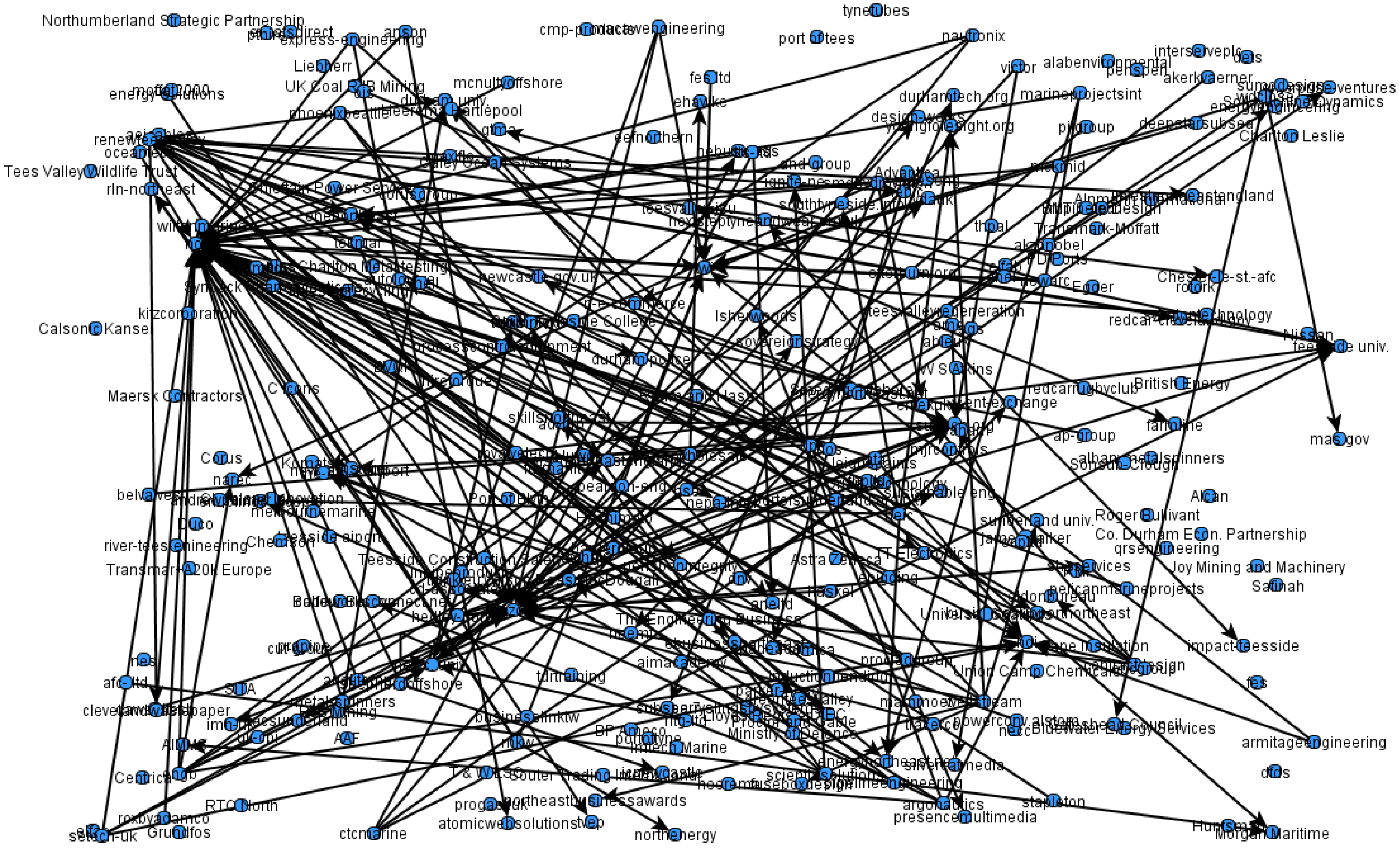


Figure 28 - Subsea: In-links (Circular network)

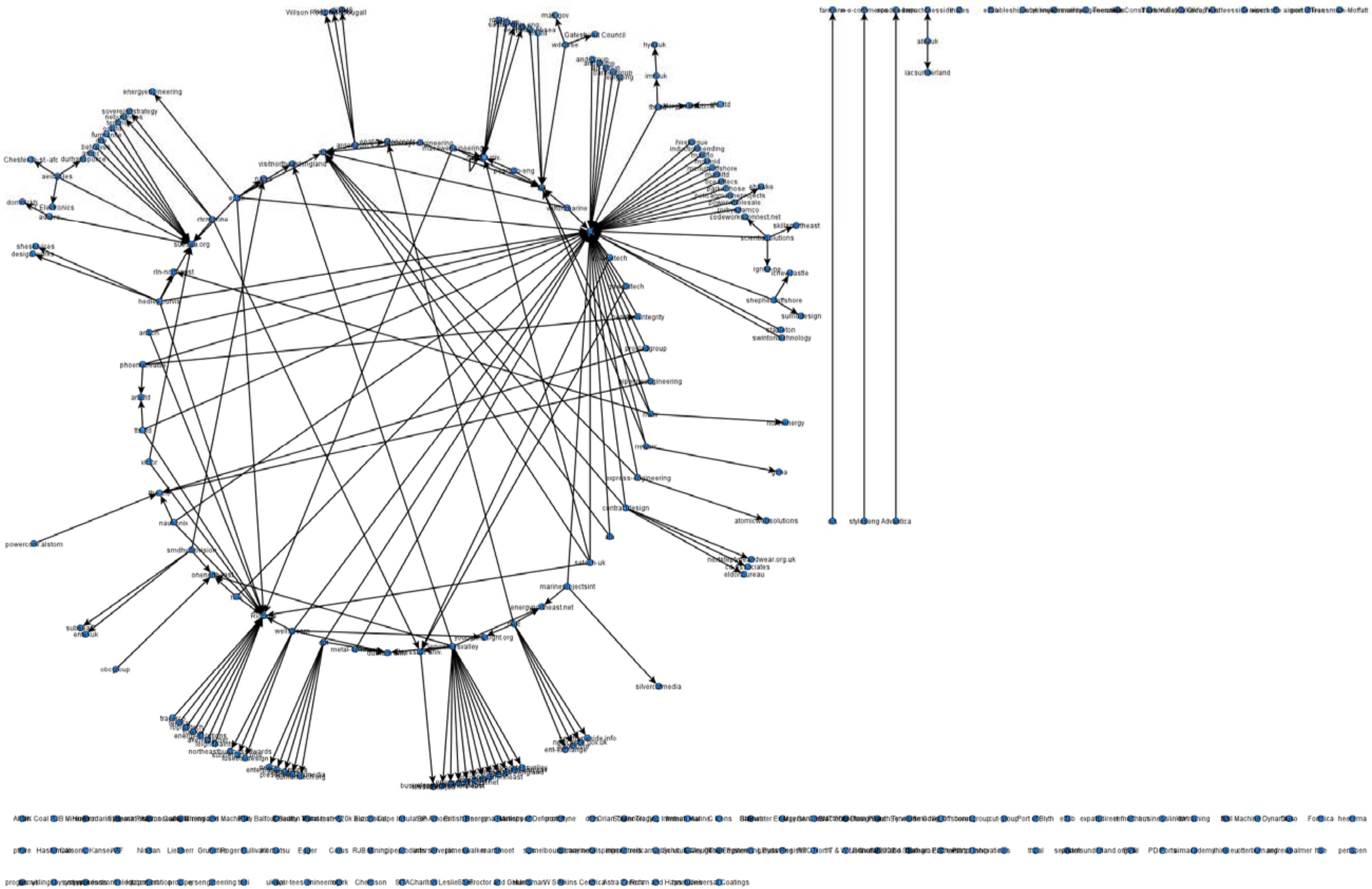
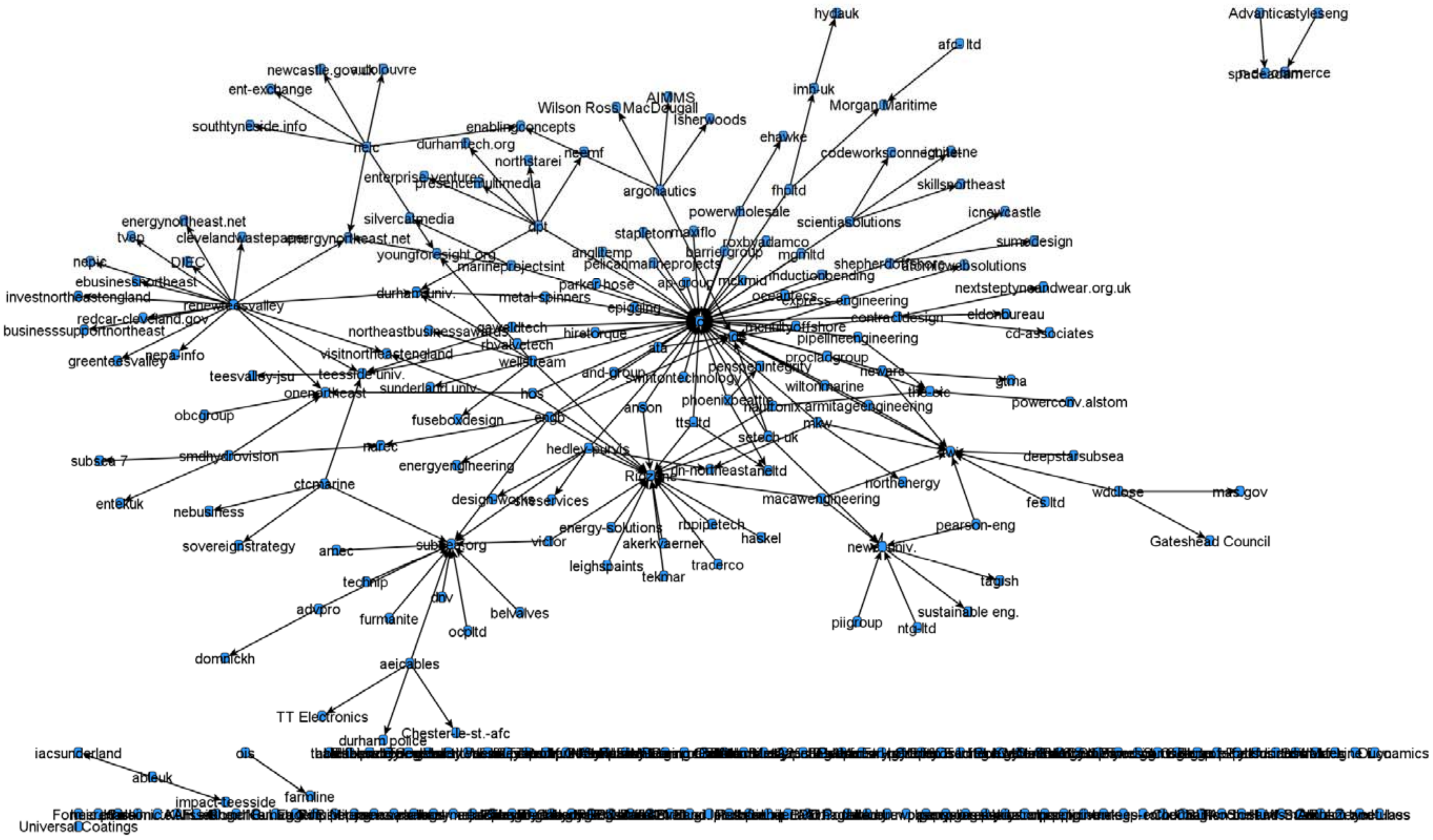


Figure 29 - Subsea: In-links (Uniform network)



11.1.6 Combined linkage set for the Subsea Cohort

The combined set of linkages brings together the three which have been the subject of examination thus far. Before merging these sets of embedded and observed links with in-links an opportunity was taken to rationalise some of the nodes and their associated linkages. In particular it was observed that in the case of some universities, when acquiring links these pointed to a particular department, school or business entity which usually had a specific URL even though it contained the university domain name. It was felt that these separate divisions within a single institution were an unnecessary complication and also would obscure the role that a university as a whole might play in the network. These separate URLs were therefore merged into a single URL appropriate to the subject institution.

A second simplification was carried out with respect to organisation names. In the work thus far URLs have been used starting with the RBKS which uses URLs to identify organisations and all the subsequent spidering work which similarly uses URLs as the base starting point. This had been carried forward even to the extent that when additional nodes were Siedlock and Andriani data or were otherwise found by inspection the additional names were subsequently replaced by their URLs. Whilst having URLs of all the organisations being looked at is useful for checking the veracity or otherwise of a found organisation there is no need for such cumbersome descriptors when drawing network graphs. It can be seen from the preceding Figures that writing a full (and sometimes long URL) onto the graph at the node leads to a considerable amount of clutter and mostly obscures nodes with a lot of links that are close together. This has occurred with 238 nodes and up to 232 links so it was felt that a merging of two networks about this size each would make the problem worse. The URLs were therefore replaced with a short node name sufficient to identify the organisation.

In combining the two sets of linkages it was important, as has been noted previously, to have the links ‘directed’ or in other words to have the out-links and in-links pointing in the opposite directions. In VISONE this is controlled by the basic input data from the .csv files. If this not done correctly, although the topography of the linkages looks the same, the richness of connections from and to nodes is diminished and metrics such as in-degree and out-degree cannot be used for comparison purposes.

Figure 30 - Combined linkages (Random graph layout)

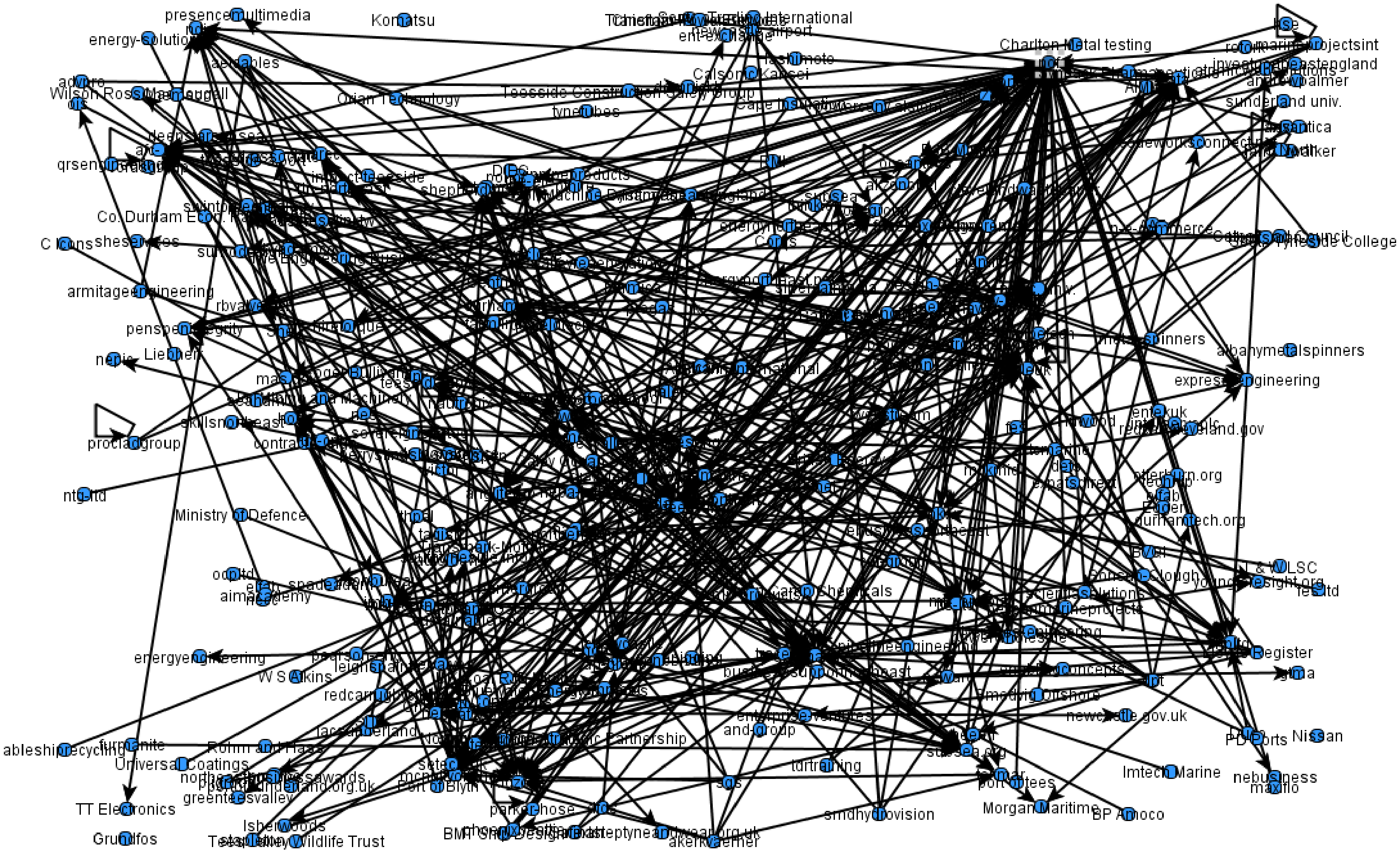


Figure 31 - Combined linkages (Circular graph layout)

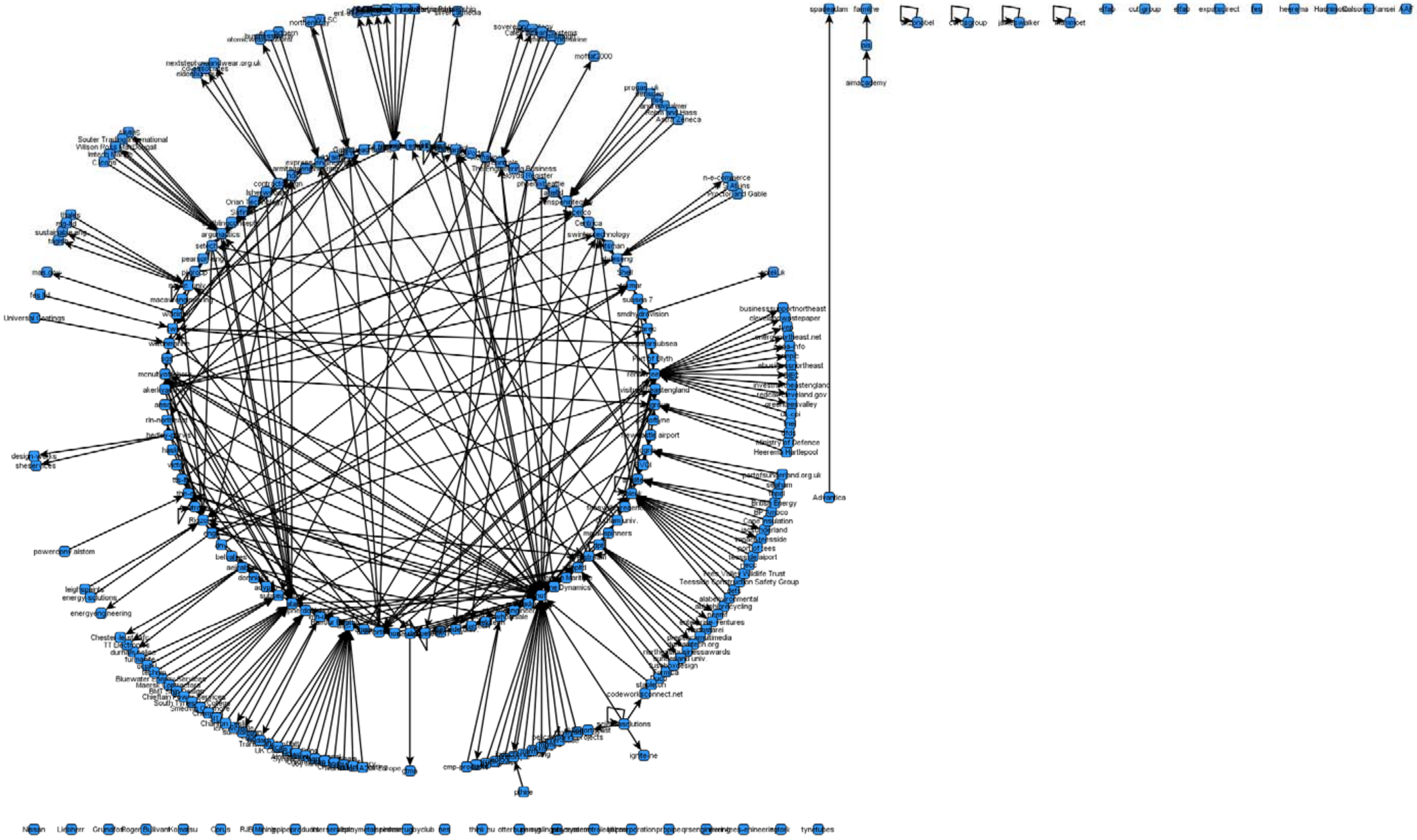
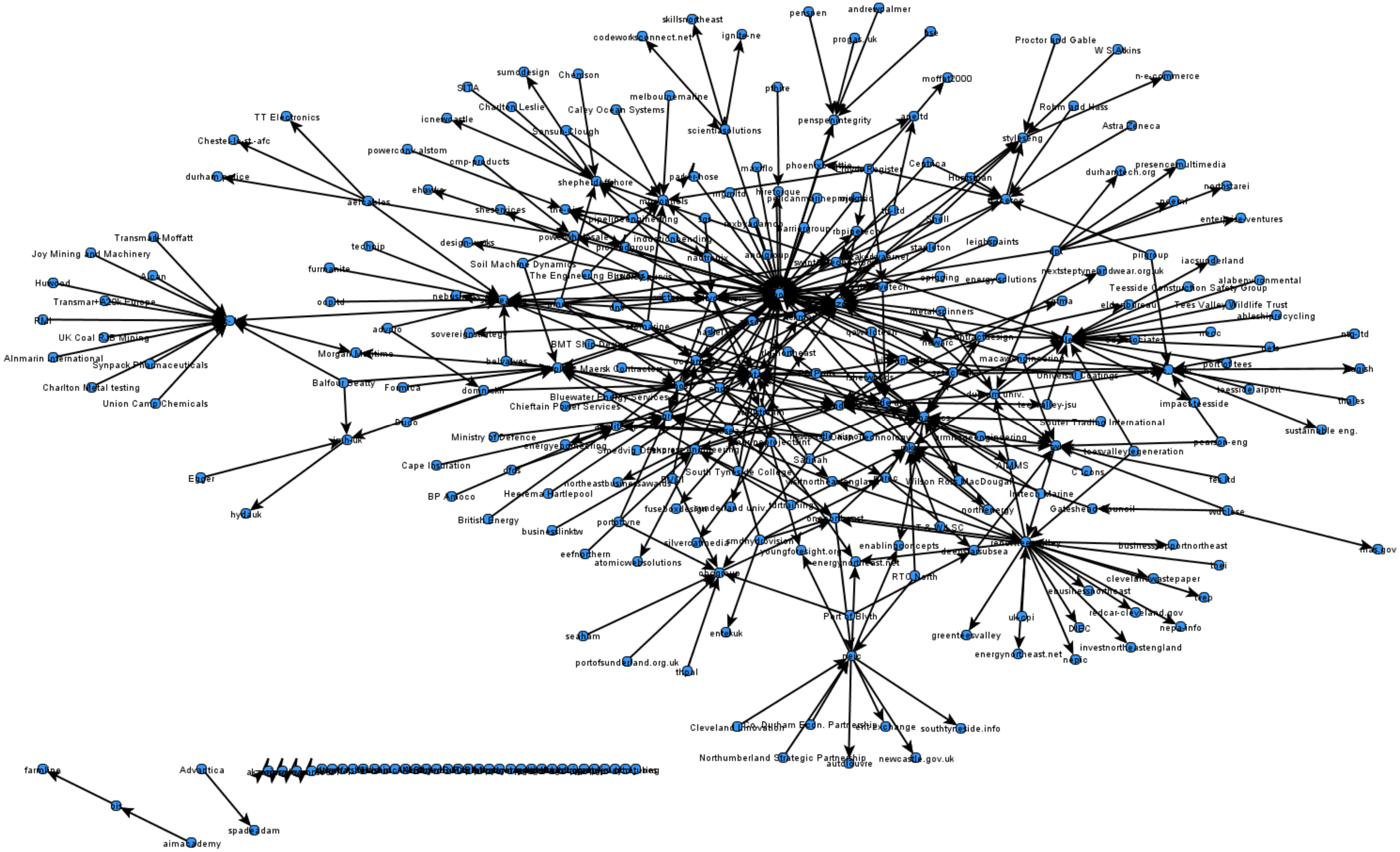


Figure 32 - Combined linkages (Uniform graph layout)



11.1.7 Results and Observations

The combined dataset for the subsea cohort contained 282 nodes and 394 links. Of these there were 32 nodes that were isolates and did not, as far as observed or derived connections were concerned link to any other. In addition there were 3 that linked only within their own website and one was an independent diode.

As a reminder these data were derived from the following sources:

1. Organisations found from the RBKS that had the word 'subsea' somewhere on their organisation's website. {nodes}
2. The subsea dataset as determined by Siedlock and Andriani. {nodes}
3. Embedded references to other nodes within the same bounded data. {links}
4. Observed references to other organisations found from #1 and #2 above. {nodes and links}
5. References from external sources on the internet that pointed to nodes in the database. {nodes and links}

The combined graphs, Figure 30, Figure 32 and Figure 32 show the very dense network generated particularly around what appear to be some key nodes (certainly as far as the links found by the methods already described are concerned). The metrics associated with the network using the number of links by in-degree and out-degree are shown in Table 10.

Table 10 - Centrality ranking for subsea links (all)

	Centrality			
Node (ID)	**Degree**	**In**	**Out**	**Comments**
nof	55	44	11	membership organisation
rigzone	22	6	16	offshore industry website
argonautics	20	13	7	private sector 'cluster' group
ableuk	18	14	4	private sector
ata	18	15	3	private sector, now part of Babcock
Rigzone	17	17	0	offshore industry website
afc- ltd	14	13	1	private sector
newc. univ.	12	8	4	educational institution
subsea.org	12	12	0	subsea industry website
ap-group	12	8	4	private sector
mcnultyoffshore	12	6	6	private sector
neic	12	5	7	public/private sector
fhpltd	11	8	3	private sector
mkw	11	7	4	private sector
wellstream	11	2	9	private sector
ndi	10	8	2	membership organisation
twi	10	9	1	private sector/membership org.
express-engineering	9	5	4	private sector
shepherdoffshore	9	6	3	private sector
tracerco	9	7	2	private sector
onenortheast	8	4	4	Regional Development Agency
hos	8	5	3	private sector

It is clear that 'nof', the Northern Offshore Federation now rebranded as NOF Energy is a key node as far as the subsea cohort is concerned. When the sites were being spidered initially, after some considerable thought and discussion with industry participants, membership sites such as NOF were treated as a special case. The reasoning behind this decision was that if the entire membership list of any such site were to be added to the links found by the spider by traversing the database feeding the RBKS it could distort the representation of any particular sector. In other words if say the offshore sector represented a relatively small percentage of the RBKS derived regional stock of businesses the addition of 300+ members could substantially increase the proportion that would be found without the benefit of the additional offshore related businesses. Such membership lists would however be particularly useful if trying to expand the network maps just associated with the interests of members.

It can also be seen that other entities that are arguably 'related and supporting' appear on Table 10 ahead of many private sector firms although what may or may not rank as such in this context is discussed further in the Conclusions.

Of particular importance are the in-links in the same way that search engines pagerank sites that have many links into them. For an organisation whether it is a company or a non traded dependency it is relatively easy to place many outgoing links on their site but unless a specialist marketing company is engaged to do so it is much less easy to acquire in-links as a measure of esteem from external organisations. In this respect the industry websites www.rigzone.com and www.subsea.org have no outgoing links but are referenced entirely by external websites. This is shown in Table 11.

Table 11 - Centrality ranking for subsea (in-degree)

	Centrality			
Node (ID)	**Degree**	**In**	**Out**	**Comments**
nof	55	44	11	membership organisation
Rigzone	17	17	0	offshore industry website
ata	18	15	3	private sector, now part of Babcock
ableuk	18	14	4	private sector
argonautics	20	13	7	private sector 'cluster' group
afc- ltd	14	13	1	private sector
subsea.org	12	12	0	subsea industry website
twi	10	9	1	private sector/membership org.
newc. univ.	12	8	4	educational institution
ap-group	12	8	4	private sector
fhpltd	11	8	3	private sector
ndi	10	8	2	membership organisation
mjrcontrols	8	8	0	private sector
mkw	11	7	4	private sector
tracerco	9	7	2	private sector
obcgroup	8	7	1	private sector
oceantecs	8	7	1	private sector
renewteesvalley	22	6	16	public sector
mcnultyoffshore	12	6	6	private sector
shepherdoffshore	9	6	3	private sector
styleseng	7	6	1	private sector

11.1.8 Comparison with earlier work on a Subsea 'cluster'

It was noted firstly in Chapter 10 that research carried out by Siedlock and Andriani (2006) had tracked the emergence of geographically bounded regional activity in subsea technologies. Part of the output of that exercise was a network diagram of the connections between the identified firms. Their data had been used to draw a network of firm associations, reflecting either supply chain or collaborative links with a network map drawn using UCINET. As these basic data were made available for this research, in order to give a comparison with firm associations augmented by web based means as developed here, the same data were used but as an input to VISONE.

The basic data gave the network map as shown in Figure 33 using the connections for collaboration although the work done by the authors looked at the relationships between firms in the cohort on the basis of 103 firms being involved in:

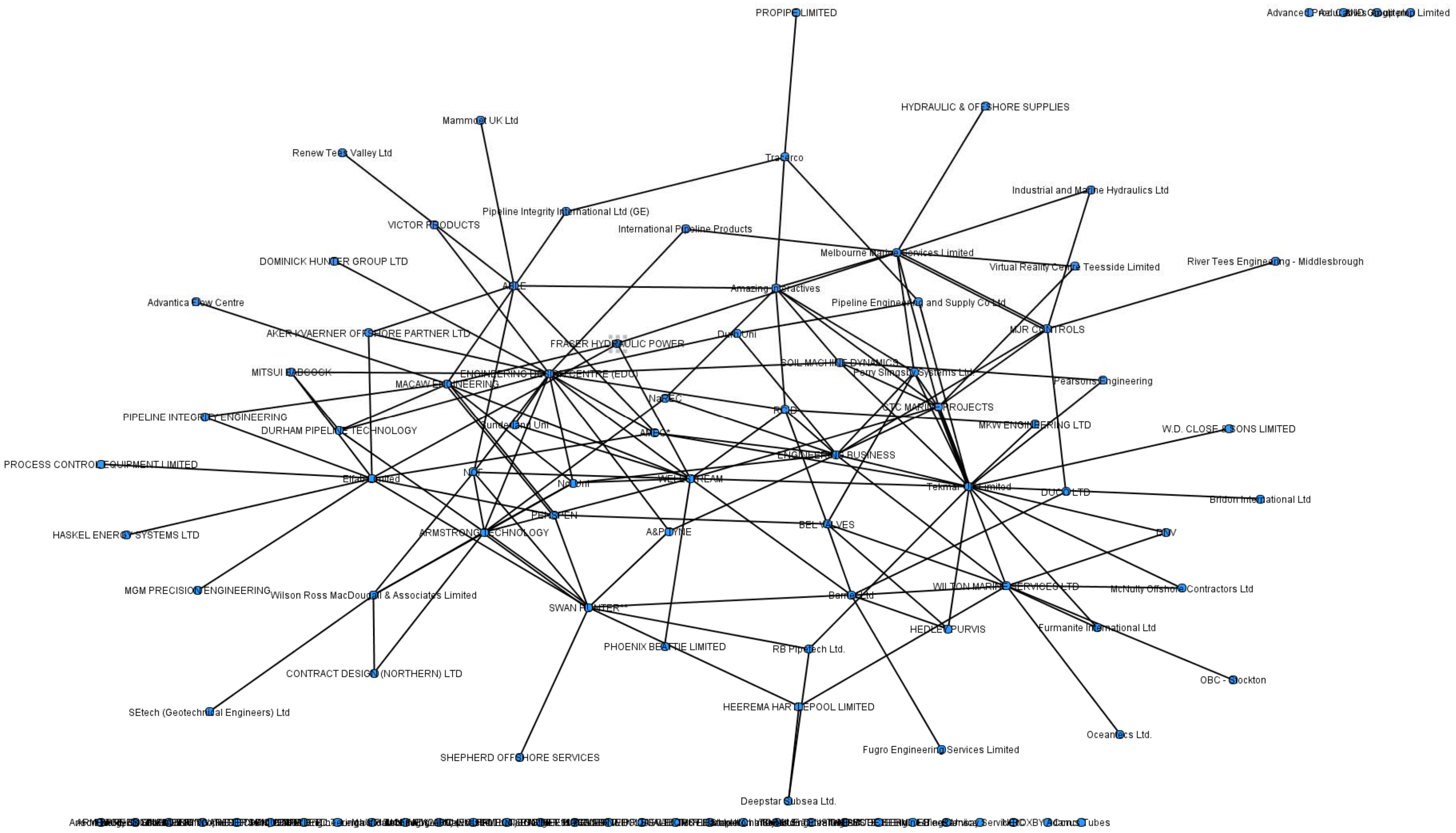
- Collaboration - 135 links
- Supplier - 154 links
- Customer – 116 links

There are a number of observations here in that as expected a number of key players emerge and as the firms were named it was therefore possible to undertake a direct comparison of who links to whom. The network is not as dense as the full combined out-links and in-links network featured in Figure 32 with the network here showing 103 nodes and 135 links. However this is an unfair comparison as part of the database (being those 80 firms for which a URL could be determined as explained on page 211), was used to augment all the additional organisations found earlier by web related means. As with the much bigger database there are a number of isolates who do not appear to connect with any other and also a number of subgroups with their own subnets of what appear to be suppliers (or of knowledge in the sense of being collaborators). As a note on the detail, to maintain a degree of consistency between the two datasets some known changes were catered for. Swan Hunter Shipbuilders had ceased trading at the time of the web based search and although still referenced on some in-links it was removed from both the new and older databases of companies.

In the web derived data different departments of a single University had been merged for simplicity. The same exercise was undertaken for the Siedlock and Andriani data for Newcastle University in that the links and nodes for the Regional Centre for Innovation in Design (RCID), the Engineering Design Centre and Newcastle University were merged into ‘Newc. Univ’.

There are, as noted above three possible networks of what are effectively collaborators, buyers and suppliers although here the one for collaborators is shown in Figure 33. In each the authors were principally interested in the topography of the network and not the direction of information flow which was inferred by the nature of the network i.e. a buyer buys from a supplier or in the case of collaboration the flow of information and knowledge works both ways. As a result of this the links are not in any way directed. It is therefore difficult to make direct comparisons as it is not known which nodes were suppliers or receivers of information or goods. This feature is of course inferred in web derived links in that firms do indicate who are their clients, supporting organisations or suppliers and the links can be directed accordingly when drawing the network.

Figure 33 - Subsea Collaborative Network (after Siedlock & Andriani, 2006)


PROPIPE LIMITED
HYDRAULIC & OFFSHORE SUPPLIES
Mammoet UK Ltd
Renew Tees Valley Ltd
Tracerco
Industrial and Marine Hydraulics Ltd
Pipeline Integrity International Ltd (GE)
VICTOR PRODUCTS
International Pipeline Products
Melbourne Marine Services Limited
DOMINICK HUNTER GROUP LTD
Virtual Reality Centre Teesside Limited
River Tees Engineering - Middlesbrough
ABLE
Amazing Interactives
Advantica Flow Centre
Pipeline Engineering and Supply Co Ltd
AKER KVAERNER OFFSHORE PARTNER LTD
MJR CONTROLS
FRASER HYDRAULIC POWER
Durh Uni
SOIL MACHINE DYNAMICS
MITSUI BABCOCK
Perry Slingsby Systems Ltd
ENGINEERING DESIGN CENTRE (EDC)
Pearsons Engineering
MACAW ENGINEERING
NaREC
PIPELINE INTEGRITY ENGINEERING
R&D
CTC MARINE PROJECTS
DURHAM PIPELINE TECHNOLOGY
Sunderland Uni
MKW ENGINEERING LTD
AMEC*
W.D. CLOSE & SONS LIMITED
ENGINEERING BUSINESS
PROCESS CONTROL EQUIPMENT LIMITED
Elfab Limited
NOF
Ncl Uni
WELLSTREAM
Tekmar Ltd Limited
DUCO LTD
Bridon International Ltd
PENSPEN
HASKEL ENERGY SYSTEMS LTD
ARMSTRONG TECHNOLOGY
A&P TYNE
BEL VALVES
DNV
MGM PRECISION ENGINEERING
Wilson Ross MacDougall & Associates Limited
WILTON MARINE SERVICES LTD
McNulty Offshore Contractors Ltd
SWAN HUNTER**
Barnet Ltd
HEDLEY PURVIS
Furmanite International Ltd
PHOENIX BEATTIE LIMITED
RB Pipetech Ltd.
CONTRACT DESIGN (NORTHERN) LTD
OBC - Stockton
HEEREMA HARTLEPOOL LIMITED
SEtech (Geotechnical Engineers) Ltd
Oceantecs Ltd.
Fugro Engineering Services Limited
SHEPHERD OFFSHORE SERVICES
Deepstar Subsea Ltd.
Tubes

The authors also reported that just 19 companies in the supply chain reported 240 linkages with 127 of these being supplier links and 116 being customer links although the figures for the full group of 103 companies are 154 suppliers and 116 customers respectively. The average for this top group was therefore over 12 links per node. The comparative figures for the web derived group as shown in Table 10 of the top 20 nodes was 14.5 links per node. Again such comparators should be used with caution. In the web derived group there is a long tail of companies with no links or with few links and the presence of a few nodes with a very large number of links does of course raise the average when just looking only at the top performers (at least from a networking viewpoint). Using web derived links does seem to have an ability to pull in links 'on the margin' i.e. firms whose association with another might not be regarded as mainstream either because their contribution might be deemed to be small or to infrequent. It may be that when doing manual snowball sampling relying on expert input the firm being questioned may not rate such firms highly in the overall supply chain or may have simply forgotten about them. The supplying firm on the other hand may feel that their contribution is worthwhile and has put an appropriate link on their website. This phenomenon was discussed in the context of a very large firm such as Shell and its numerous small suppliers to a Shell brokered project. So there will be expected differences in outcome between the manual methods of discerning linkages and the more 'hands off' approach adopted here.

The nodes identified in the collaborative group by manual means as being 'important' or 'keystones' or 'authorities' in a network sense were also compared with those identified by web based methods. This comparison is shown in Table 12. To construct the table the top 50 nodes by degree centrality in the Siedlock and Andriani work using manual methods is shown in the first half (columns) of the table and highlighted in red. The second half of the table refers to those organisations found by web based means. Where the same firms occur in the first part of the table they are also highlighted in red. For the collaborative network there are therefore some 19 of these common organisations or 38% in the top 50 displaying linkages found entirely off the web.

Table 12 - Comparison of authoritative nodes (Top 50) by Collaboration

Top Collaborative links by Manual observation				Top links from Web	
Company	**Degree**	**Comments**		**Company**	**Degree**
Newcastle Univ.	28	educational institution		nof	55
Tekmar UK Limited	17	private sector		renewteesvalley	22
WELLSTREAM	12	private sector		argonautics	20
CTC MARINE PROJECTS	11	private sector		ableuk	18
ENGINEERING BUSINESS	11	private sector		ata	18
Melbourne Marine Services Limited	10	private sector		Rigzone	17
WILTON MARINE SERVICES LTD	10	private sector		afc- ltd	14
Elfab Limited	9	private sector		newc. univ.	12
MACAW ENGINEERING	9	private sector		subsea.org	12
ARMSTRONG TECHNOLOGY	8	private sector		ap-group	12
Amazing interactives	8	private sector		mcnultyoffshore	12
MJR CONTROLS	7	private sector		neic	12
ABLE	6	private sector		fhpltd	11
Barrier Ltd	6	private sector		mkw	11
BEL VALVES	6	private sector		wellstream	11
Perry Slingsby Systems Ltd:	6	private sector		ndi	10
AMEC*	5	private sector		twi	10
SOIL MACHINE DYNAMICS	5	private sector		express-engineering	9
WRM	5	private sector		shepherdoffshore	9
NOF	5	membership organisation		tracerco	9
DURHAM PIPELINE TECHNOLOGY	4	private sector		onenortheast	8
FRASER HYDRAULIC POWER	4	private sector		hos	8
PENSPEN	4	private sector		mjrcontrols	8
Tracerco	4	private sector		obcgroup	8
AKER KVAERNER OFFSHORE PARTNER LTD	3	private sector		oceantecs	8

HEDLEY PURVIS	3	private sector		setech-uk	8
MITSUI BABCOCK	3	private sector		durham univ.	7
Pipeline Engineering and Supply Co Ltd	3	private sector		amec	7
NaREC	3	public/private sector		contractdesign	7
Durh Uni	3	Educational institution		dpt	7
A&P TYNE	2	private sector		engb	7
CONTRACT DESIGN (NORTHERN) LTD	2	private sector		styleseng	7
Deepstar Subsea Ltd.	2	private sector		swintontechnology	7
DNV	2	Classification society		akerkvaerner	6
DUCO LTD	2	private sector		anglitemp	6
Furmanite International Ltd	2	private sector		ctcmarine	6
HEEREMA HARTLEPOOL LIMITED	2	private sector		hedley-purvis	6
Industrial and Marine Hydraulics Ltd	2	private sector		marineprojectsint	6
International Pipeline Products	2	private sector		Lloyds Register	6
McNulty Offshore Contractors Ltd	2	private sector		newarc	6
MKW ENGINEERING LTD	2	private sector		penspenintegrity	6
PIPELINE INTEGRITY ENGINEERING	2	private sector		scientiasolutions	6
Pipeline Integrity International Ltd (GE)	2	private sector		wiltonmarine	6
RB Pipetech Ltd.	2	private sector		aeicables	5
Virtual Reality Centre Teesside Limited	2	Educational institution		narec	5
Sunderland Uni	2	Educational institution		the-eic	5
Pearsons Engineering	2	private sector		nautronix	5
Advantica Flow Centre	1	private sector		powerwholesale	5
Bridon International Ltd	1	private sector		tekmar	5
DOMINICK HUNTER GROUP LTD	1	private sector		teesside univ.	5
50				19	

In view of the possibility of more organisations common to both lists appearing on the connections found by Siedlock and Andriani their data was sorted such that the top links could be found for the ‘Suppliers’ and for the ‘customers’ as well as the ‘collaborators shown above.

Each of these three segments was looked individually and in addition they have been merged and ranked by total links. Some firms of course appear in all 3 lists but generally those who are collaborators are not customers and those who are suppliers are not customers although this is by no means a universal rule. Table 12 - Comparison of authoritative nodes (Top 50) was thus expanded as matching nodes from the ranked Siedlock and Andriani data were added (for the top 50). The net result of this link counting process was as shown in Table 12. Additional matching companies picked up by web based methods that occur on one or more of the Siedlock and Andriani lists bring the total to 22 companies out of the top 50 being a 44% match. Whilst this may seem either quite reasonable or low dependent on expectation it is important to bear in mind the sources of the acquired linkages. There are a number of organisations that were acquired by web based means that would be unlikely to feature on a list of mainstream suppliers if say a Purchasing Manager in a subsea machinery constructor was queried regarding his or her purchasing requirements. Such organisations may be below the radar of company management preoccupied with purchasing decisions for all their key materiel and information but nevertheless support the subject industries in their own way. In this case in Table 13 the following nodes could fall into this category and are discussed briefly below.

Subsea.org and rigzone.com are web based information sources included for reasons discussed when deciding which category of organisation to include as initial search criteria. Argonautics is an organisation, not a single company and it would not be a contractor although many of its constituent members such as ATA, WRM and SeTech do appear. Renewteesvalley is a public sector organisation charged with promoting a variety of regeneration projects including industrial renaissance in a specific sub-region. NDI (Northern Defence Industries) is a trade and lobbying group set up to promote defence related suppliers in the region. As such many firms who are part of the subsea and offshore segment also sell into defence related markets (this is a theme that is taken up in the main conclusions). However in the context of subsea, again it is an adjunct rather that a primary supplier or collaborator.

TWI, is another membership organisation (although it also undertakes paid-for consultancy) in the areas of materials science. It would be expected that it would feature in materials specification and even research but it may be that companies pay their membership and regard it in a similar way to the Chamber of Commerce in that it is there in the background if required to help solve problems and in any case some assistance is included in the annual membership fee.

If we were to take out the above 6 organisations on the basis that, for comparative purpose they are unlikely to figure on the list of direct suppliers, customers or collaborators then the match with web derived compared with Siedlock and Andriani derived companies rises to 50%.

Table 13 - Comparison of authoritative nodes (Top 50) All sources

Top links from Web	
Company	**Degree**
nof	55
renewteesvalley	22
argonautics	20
ableuk	18
ata	18
Rigzone	17
afc- ltd	14
newc. univ.	12
subsea.org	12
ap-group	12
mcnultyoffshore	12
neic	12
fhpltd	11
mkw	11
wellstream	11
ndi	10
twi	10
express-engineering	9
shepherdoffshore	9
tracerco	9
onenortheast	8
hos	8
mjrcontrols	8
obcgroup	8
oceantecs	8
setech-uk	8

durham univ.	7
amec	7
contractdesign	7
dpt	7
engb	7
styleseng	7
swintontechnology	7
akerkvaerner	6
anglitemp	6
ctcmarine	6
hedley-purvis	6
marineprojectsint	6
Lloyds Register	6
newarc	6
penspenintegrity	6
scientiasolutions	6
wiltonmarine	6
aeicables	5
narec	5
the-eic	5
nautronix	5
powerwholesale	5
tekmar	5
teesside univ.	5

Conversely however, there are firms that rank high on one list that do not appear at all in the other and vice versa. There are a number of reasons here for these phenomena.

Firstly, as already discussed very large contractors have a propensity not to put links, embedded or overt on their sites. Thus these large firms rank low on such out-links. Whilst this effect is moderated to some degree by a larger number of smaller firms citing them and appearing as in-links to the large firms, the degree of centrality is correspondingly low when all the links are summed. This can be seen in the differences between Figure 25 and Figure 28. Although some of these links are hard to discern from the mesh what seems to be happening is that the centrality of the large companies is quite low for in-links but very small or non-existent for embedded or visible links. Examples of larger firms 'missing' or not in the top 50 of the internet derived network are Amec, Pearsons, the Engineering Business and DUCO.

By contrast the related and supporting organisations that do not appear on the Siedlock and Andriani list but which were discovered by internet searching included organisations such as subsea.org, rigzone.com, public sector economic development websites and small trade organisations together with a number of firms whose main line of business was in sectors such as automotive or mining equipment but who had expertise valued by a small number of firms in the subsea cohort.

Another issue is one of timing. The groundwork by Siedlock and Andriani was undertaken in the years 2005-6. There is thus up to a 2 year gap between this work and the data and information picked up off the web although it is acknowledged that web derived information may not always be fully up to date either. The accuracy as well as the currency of web derived data was a topic discussed in Section 9.5 in the context of spidering of text from web sites. That discussion considered how valid any company website might be for both accuracy and completeness and came to the broad conclusion that firms do tend to pay attention to their website with regard to keeping it current as for most it is a marketing tool. Thus the found websites are likely to reflect both appropriate companies not found before and also others moving into the subsea sector as that market expands. In conclusion there are probably more firms available now than 2 years ago when the first study was carried out.

A final factor to consider when comparing the two methods of discerning networks is one that has been alluded to throughout the research and that is one of sampling. The 14000 URLs used to find firms for the subsea cohort is a minority sample of all firms in the region and were it possible to obtain a full set of URLs for all regional firms then the number of firms found by this method would rise and other firms would be added to the corpus. We have of course, in this case added manually the Siedlock and Andriani cohort of firms which materially improves the 'hit rate' for finding firms involved in subsea activities before going on to use all firms in the database for evidence of some link between them. In undertaking the comparisons between the Siedlock and Andriani data sources and the shortcomings of the web derived data it should also be noted that snowball sampling used to elicit the former is by no means a perfect method. There are two major limitations and weaknesses of snowball methods [Hanneman and Riddle (2005)]. Firstly actors who are not connected i.e. isolates are not located by this method. The presence and numbers of isolates can be an important feature of of populations for some analytical purposes. The snowball method may thus tend to overstate the 'connectedness' or 'solidarity' of poulations of actors. Secondly there is no guaranteed way of finding all of the connected individuals (or organisations) in the population. If the 'snowball' is started rolling in the wrong place or places we may miss whole sub-sets of actors who are connected but not attached to the starting point.

In conclusion, when comparing the networks of firms in the subsea cohort obtained by two fundamentally different approaches and in spite of the shortcomings as discussed the similarities are obvious; the topography of the networks is similar, the connections are of the same order and the presence of 'keystone' organisations can be discerned.

Note on the use of Social Network Analysis and graphing programs

It is perhaps worth commenting on the use of these comparison metrics in this and indeed in a wider context. The academic standard program for social network analysis would seem to be UCINET with the associated visualisation done by NetDraw[32] for portraying the actual network. UCINET is particularly strong on a variety of metrics for analysing the strength of ties and various clustering co-efficients whilst Netdraw has a facility for example to manually move nodes around as if connected by 'springs' to enable clarity of observation around a particularly authoritative node. These programs are well enough known and indeed were examined for use (together with a number of others[33]) when the early networks subsequently started to look more complex. However the use of VisOne was continued up to this point for 3 reasons:

1. The program can accept as input .csv files or input directly from MS Excel. As all the link data was in this format the data transfer process was seamless and hence accurate.
2. The program was free and available
3. A number of required analysis metrics were claimed to be available including for Centrality;
    a. Degree
    b. Indegree

[32] www.analytictech.com/

[33] A part summary of Social Networking Analysis software together with associated visualisation and data manipulation programs are shown in http://vlado.fmf.uni-lj.si/vlado/vladonet.htm and as discussed briefly in Network Analysis

c. Outdegree

Only the first three above have been reported on thus far the reason for this being it was felt that the basic link data upon which the whole of the resultant networks were based was not particularly robust in a statistical sense. As a consequence of this it was felt that application of ever more esoteric metrics for network analysis might only serve to mask this fact. In other words it is known that the link data is biased because of the way links are implemented by their authors and additionally there is no observable 'strength' in any individual link. As link 'strength' can be important as an input for some network metrics, here it is not so much a case of any one link being more 'important' than any another but that there is no way of measuring it. It was felt therefore that attempts to look at the many possible network/link attributes other than for those associated with centrality might be misleading. To continue the cautionary approach the whole question of 'properties' of linkages is one that should be addressed as the term 'properties' can mean anything from the 'strong friendship' in a social networking sense to a trading relationship between firms. In Siedlock and Andriani's work the authors looked at ties on 3 bases being knowledge exchange or more usually knowledge procurement, a buyer relationship and a supplier relationship and three separate networks using mostly the same firms were established. In the diagrams discerned here by web based means the nature of the individual relationship and hence their properties are not known, only the direction on the basis of whether it is an out-link or and in-link. We may guess at the type of relationship if for example a University is referenced on a website then a knowledge exchange relationship might exist but we cannot be sure.[c.f. Granovetter M.S. (1973 pp. 1360-1380)]

There is no doubt that the 'nature' of internet derived inter-organisation linkages and their inherent properties is an area worthy of further investigation and the possibility of initially ascribing say a level of confidence to individual links would perhaps open up the possibility of use of more sophisticated network studies. This is noted in recommendations for further research.

As has been commented on many times throughout the thesis the research question is to see what can be done using a sample of firms as a trial, to develop a methodology and to understand the inherent limitations and how these limitations might be overcome. Therefore as with the work done on text mining in Chapter 4 if it were possible to work with a full set of URLs it should also be possible to establish all firms involved in the subsea segment. However with link tracing we would also require to have internet derived links as a near analogue of actual link relationships, a situation considered unlikely unless firms adopted the www.sevenrings.co.uk approach on a large scale. Nevertheless there is some correlation between linkages derived by web based means and those derived by snowball sampling. What is perhaps more interesting is that web based means can find linkages, particularly through in-links, that even the firms (or probably more accurately the individual/s within the firm) questioned in the snowball sample were not aware of the existence of. These points are discussed further in the Conclusions.

11.2 Further investigations into the nature of the Network

In the immediately preceding section comments were made about why the program Visone was used thus far, the main point being that given the quality and kind of linkages discerned it would not be appropriate to undertake an analysis of the network using metrics other than those based on centrality involving the number of links. Even with this restriction Visone's capability does have a few gaps (not all of the metrics described in the documentation have in fact been implemented in the version used) and for this reason (and also as an alternative presentation) it was decided to use UCINET to look at some metrics describing not just individual node characteristics such as Egonets but also those that describe the network in the round i.e. in a holistic way.

The same basic data as used for Figure 32 being the final subsea networks were also used as input for all subsequent analyses in this section.

11.2.1 Further examination of the Subsea Network using UCINET

UCINET has a wide range of capabilities for examining the structure of networks and bearing in mind the health warnings above regarding the application of complex metrics to doubtful data the following metrics based on straightforward nodes and directed linkages were considered for further exploration.

- Degree (using different definitions)
- Closeness
- Betweenness or flowbetweenness

- Eigenvector
- Clustering coefficients
- Egonets
- Other centrality measures

If possible a graph of the relations has been drawn but for some of the whole network metrics calculated information is more appropriate. Where not explained at the time a definition of these measures is given in Appendix 13 Definition of SNA terms.

The basic binary data used to draw the final network shown in Figure 32 was used as the matrix input for UCINET. In UCINET it is also possible to label the type of node or link according to a perceived 'membership' category and here this was undertaken with the nodes to try and discern the part played by such different categories of node particularly when looking for the presence of 'related and supporting' players in the network. The nodes were therefore labelled as follows:

1. Private companies. {colour RED}
2. Support organisations including non-profits, Universities, Societies such as Lloyds Register. {colour BLUE}
3. Other support organisations sponsored by the private sector such as Rigzone, subsea.org., Engineering Employers Federation. {colour BLACK}
4. Govt. agencies, RDAs, Local authorities. {Colour GREY}

The program Netdraw was used in conjunction with UCINET to portray the network map and the different types of node (1,2,3,4) as above can be shown either as a different colour or shape to help with identification. Here the colours as noted above were used. The overall network map is as shown in Figure 34. This a map with all the isolates and pendants shown and consequently is rather cluttered although the actual map has been drawn using a spring embedded algorithm based on geodesic distance.

The figures that follow are an attempt to dissect this network such that the most important or influential nodes are discerned. This might be important in the sense that from theory, industrial clusters often form around such nodes but it may be that in many cases such nodes, particularly on the support service side actually came later.

The network maps subsequent to Figure 34 all have the isolates and the pendants removed but the UCINET derived metrics have them included. Removing such unconnected or sparsely connected nodes does have the advantage of improving clarity of the linkages and various associated sub-networks but the UCINET metrics preserve the real world measures by including all 282 nodes. The figures and the rationale behind them are discussed next.

Figure 35 - Type 1 Firms (Private Sector)

These graphs use the ability of UCINET to ‘switch off’ certain types of node and in Figure 35 the nodes shown are linkages entirely with Type 1 firms being those normal trading firms within the private sector. As can be seen there is a significant interaction with few unconnected isolates (bearing in mind that the isolates have been removed from the starting set of binary data). It could be inferred that there are significant connections amongst this type of firm with a number of larger ones such as MacNultyOffshore and Amec showing some particular influence. This is interesting as the pure centrality counts shown in Table 10 did not show these firms as influential as some of the others

Figure 34 - Subsea network (all by type)

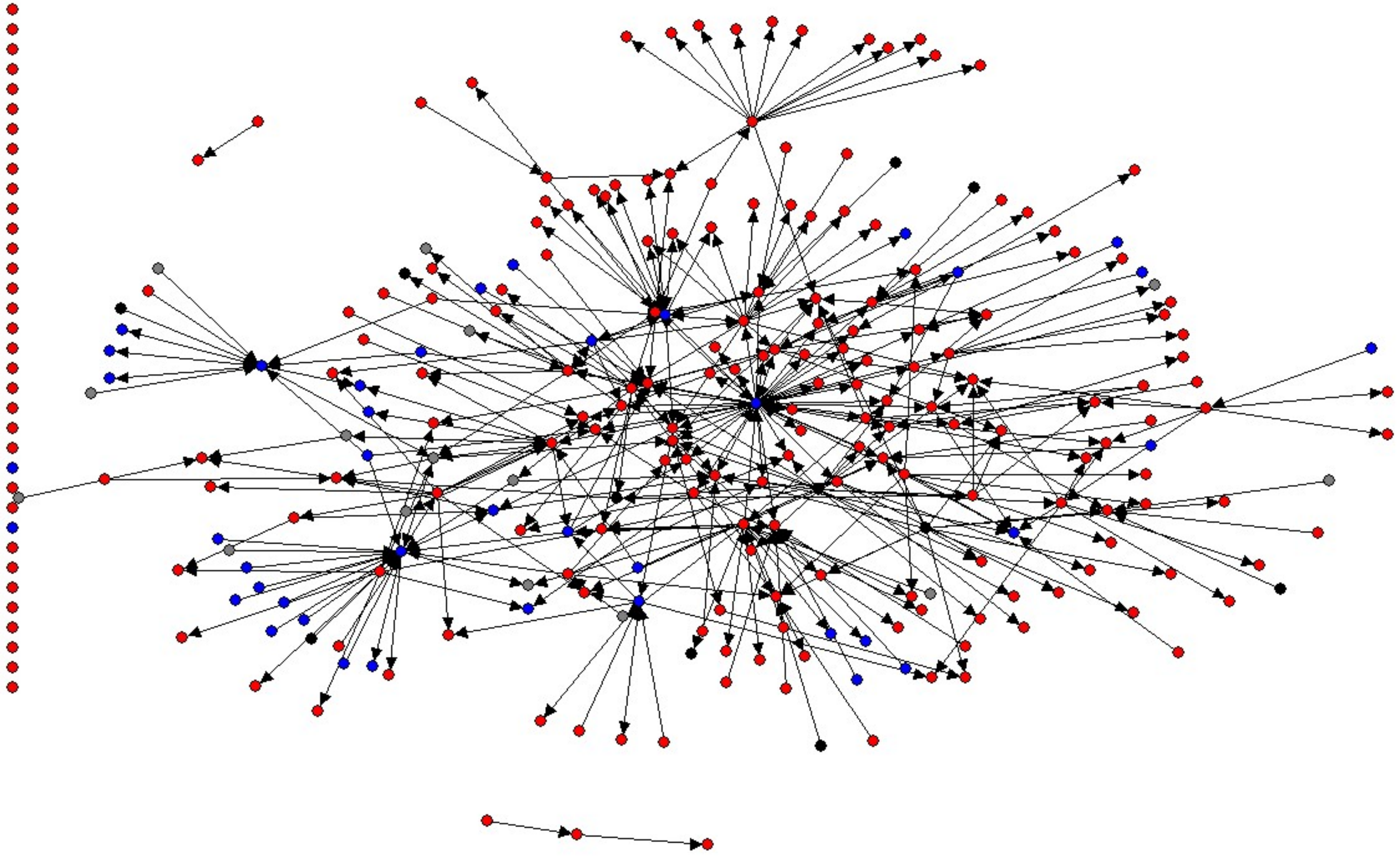


Figure 35 - Type 1 Firms (Private Sector)

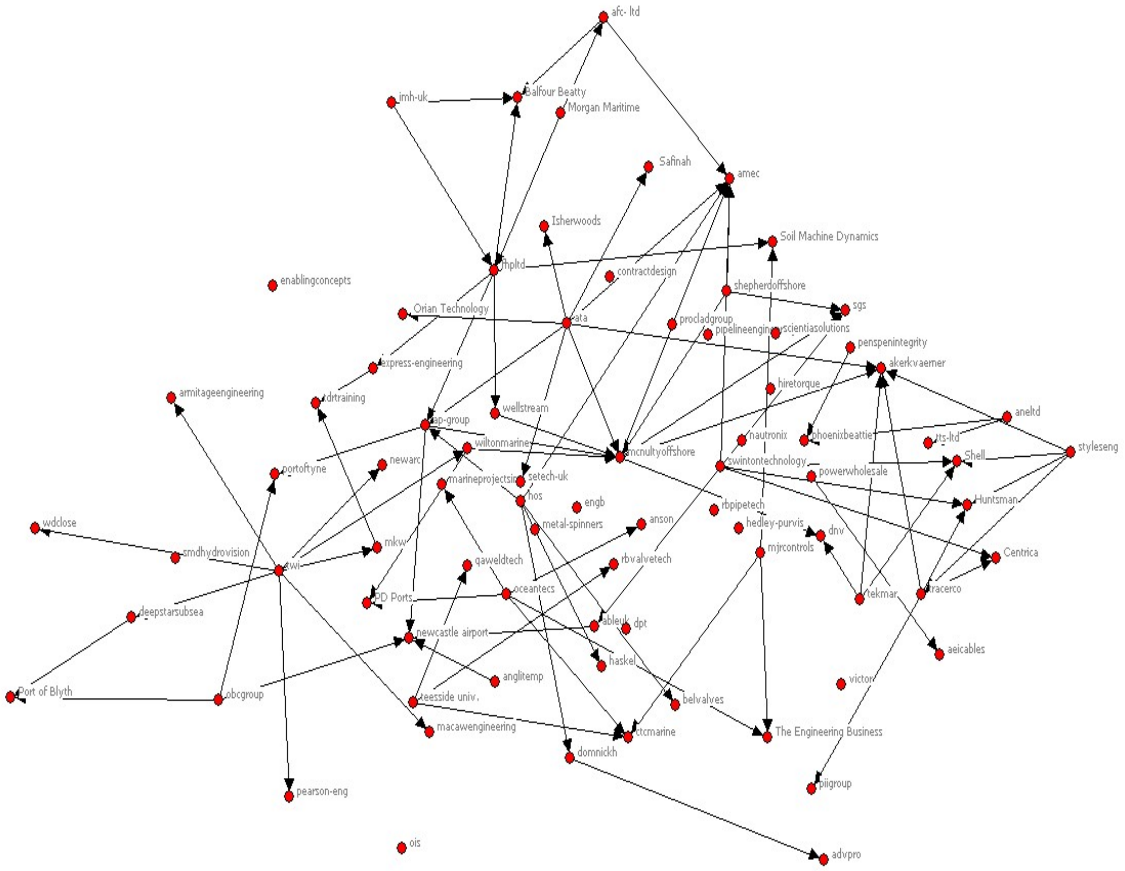


Figure 36 - Type 2,3,4 organisations

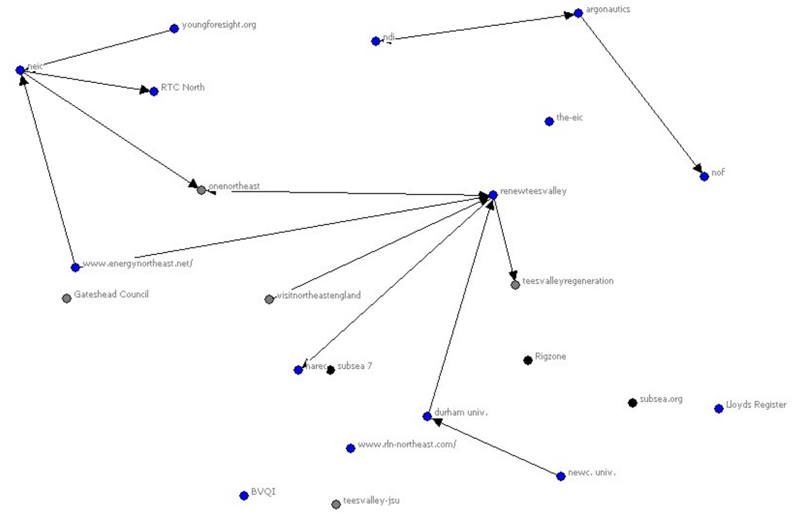


Figure 37 - Egonet; NOF

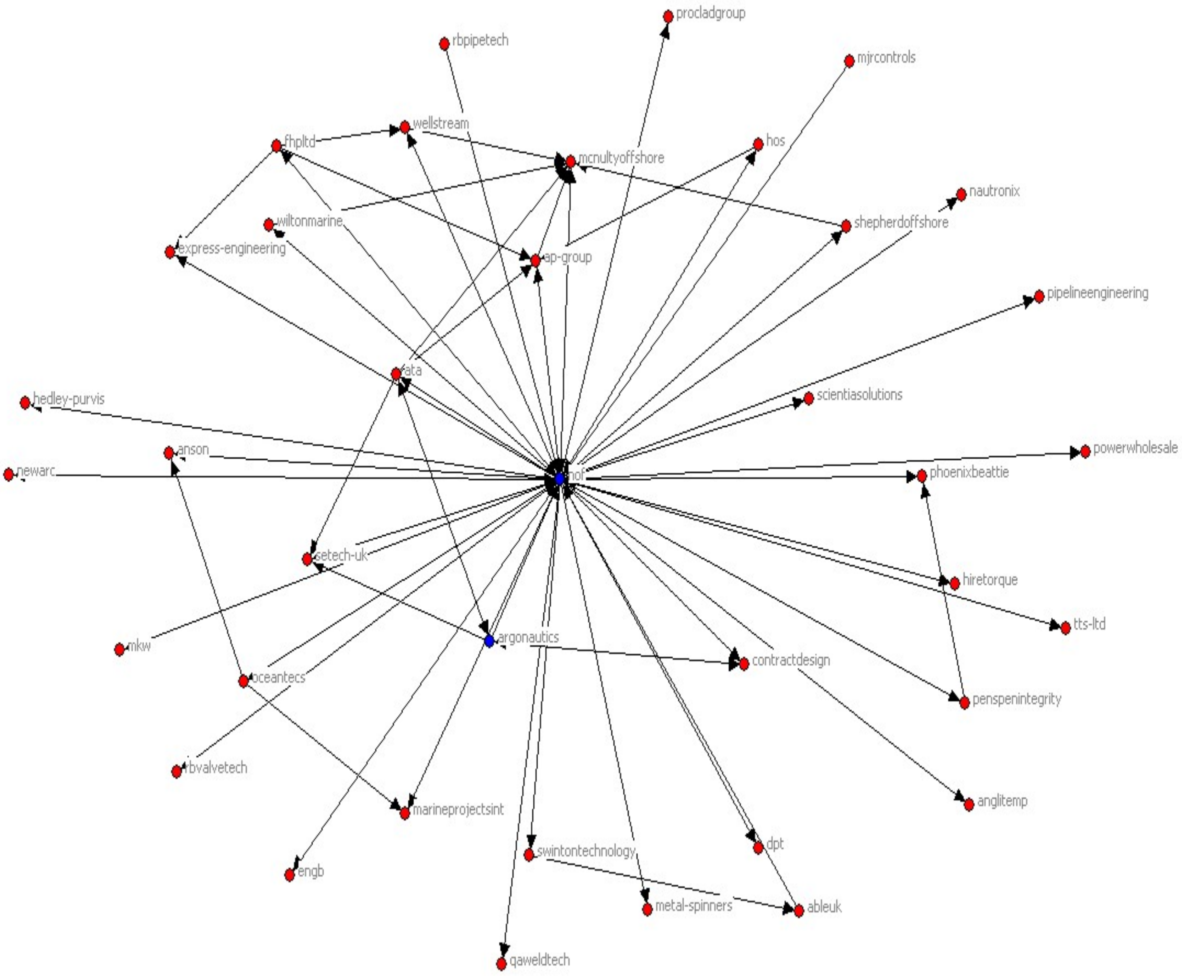


Figure 38 - Egonet; Argonautics

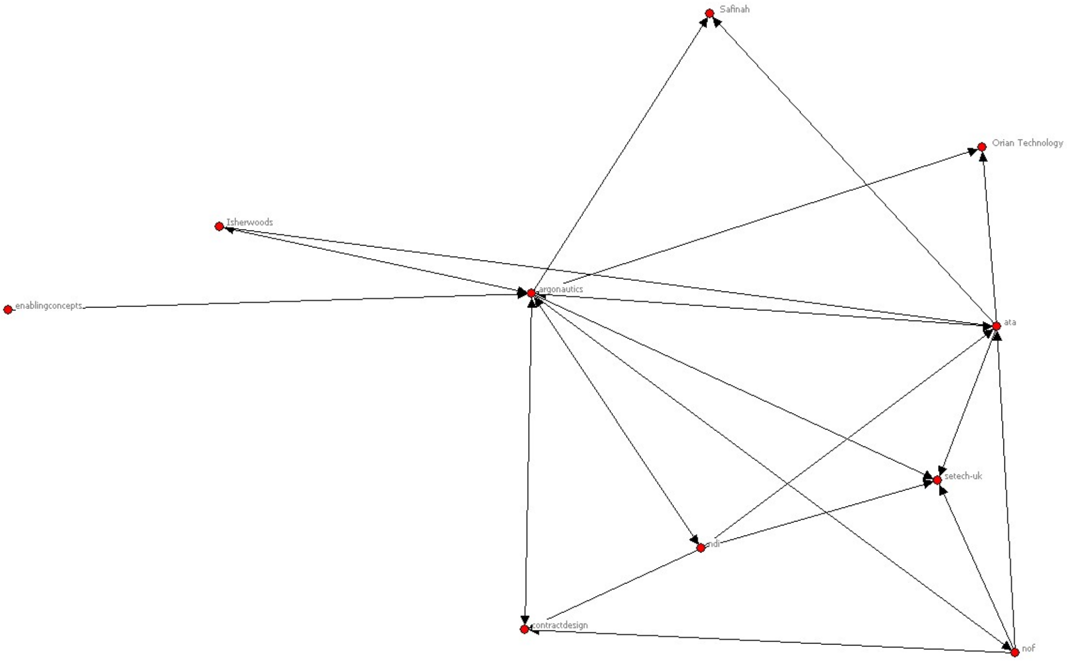


Figure 39 - Egonet; Rigzone

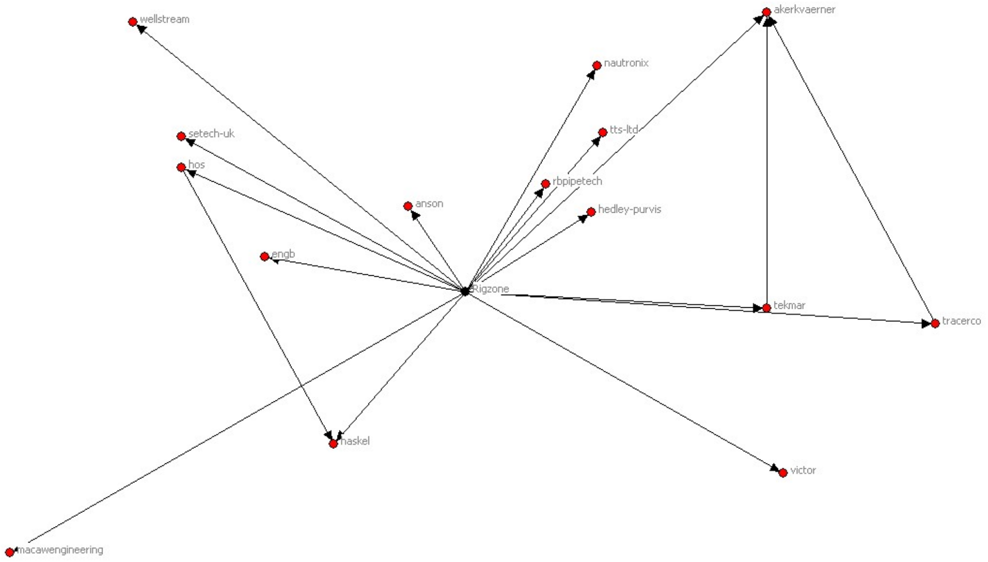


Figure 40 - Egonet; ata

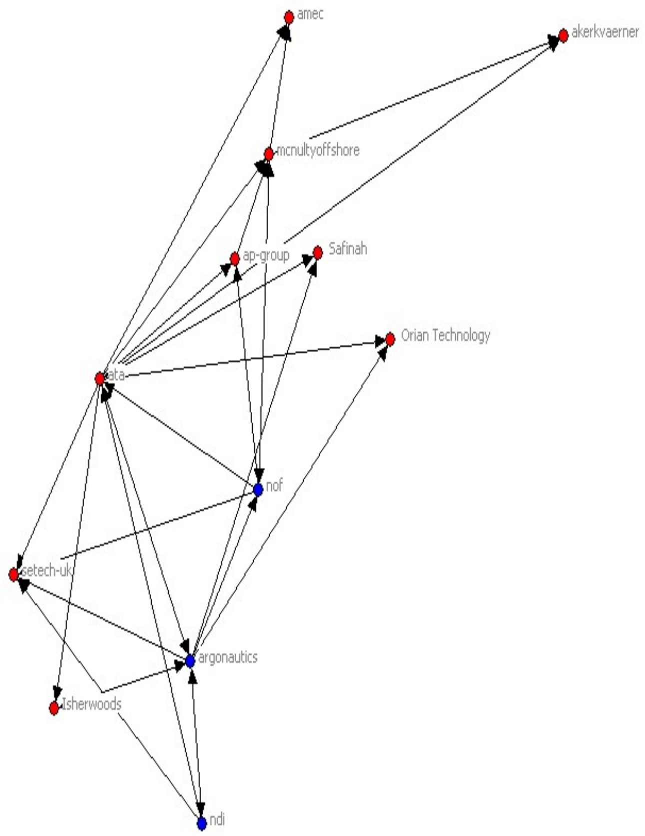


Figure 41 - HiClus of Geo Distances; NOF

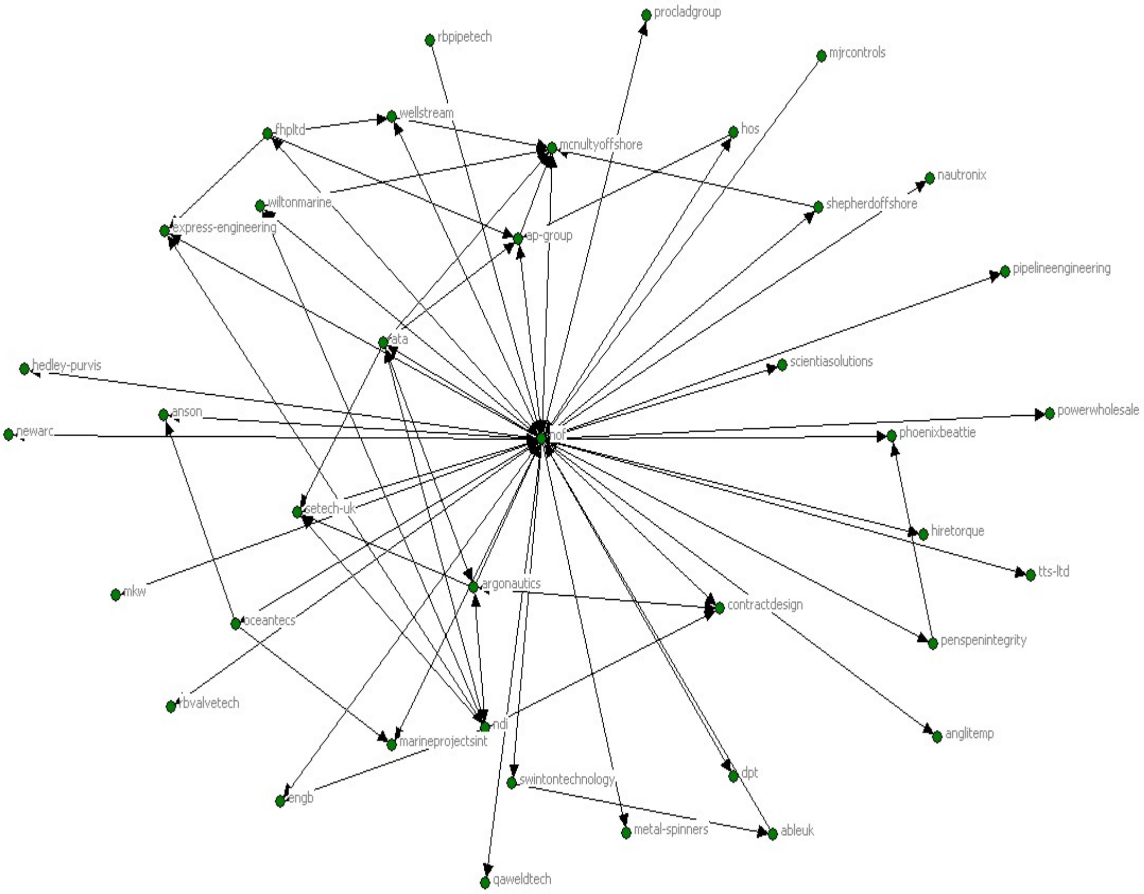


Figure 42 - HiClus of Geo Distances; RenewTeesValley

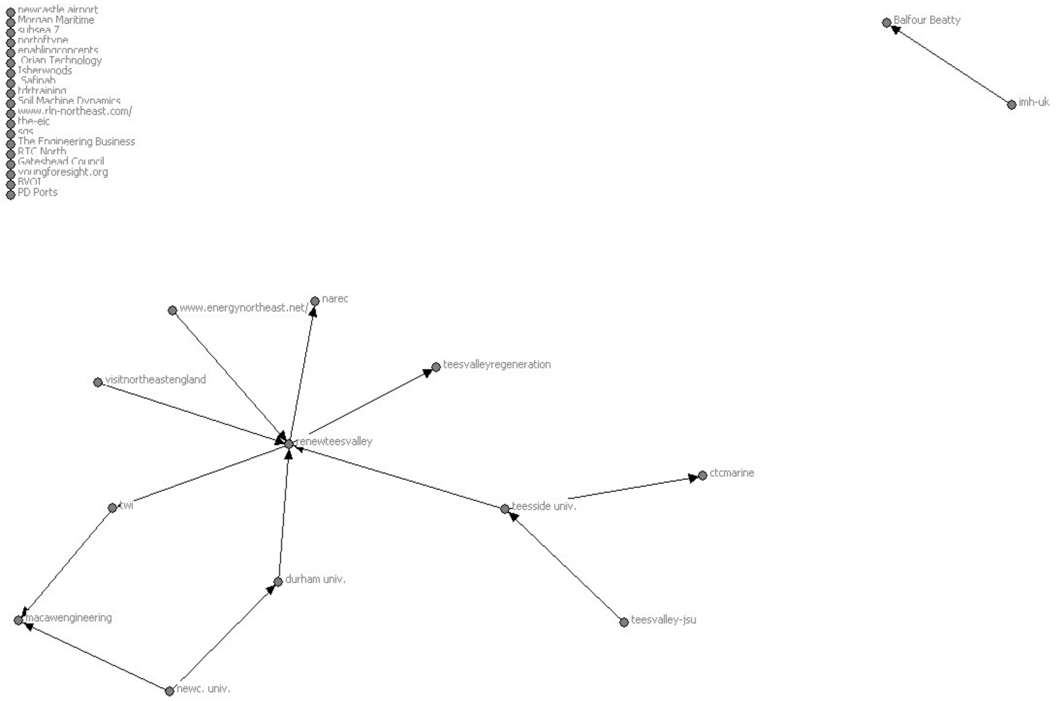


Figure 36 - Type 2,3,4 organisations

This figure shows only those organisations that fall into the categories of type 2, 3 or 4. Although they are numerically smaller than the type 1 firm they are not as well connected internally and there are a number who are not connected to their peer group at all. This not to say that they are not as well connected as the type 1 organisations as in the full graph of Figure 34 they do connect with a wide variety of private sector firms. It is just interesting that apart from renewteesvalley there are few linkages to other support organisations.

Figure 37 - Egonet; NOF

UCINET has a facility to again ‘switch on’ Egonets to explore those of interest. The next 4 charts have been looked at because of their either size of prominence in the network. The first of these is NOF which is not surprising as there are a number of other measures such as centrality in which this organisation features prominently. A feature of egonets is that here they show the direction of arrows as out from the ‘ego’ towards its ‘alter’ which is the opposite of most of the directions for the NOF in the overall network. There are also a number of subnets amongst the ‘alters’. This is the largest egonet in the overall network and indicates the high degree of influence or ‘power’ that the NOF has.

Figure 38 - Egonet; Argonautics

Argonautics is a small grouping of marine firms that has been noted at various points in the text. Here we can see that although not as dense as the NOF it still has an egonet based mostly on its constituent members.

Figure 39 - Egonet; Rigzone

Rigzone was chosen as it is one of the few type 3 firms and all its links are in-links. As a result of this it has an egonet of alters composed entirely of mainstream private sector firms.

Figure 40 - Egonet; ata

ATA is a private sector firm active in engineering design in a number of markets. It has alters in both the private sector and also a number of support organisations including Argonautics of which it was a founding member. The support organisations are the NOF, Argonautics and Northern Defence Industries (ndi). ATA could be regarding as a good middle of the road example of different types of links and this shows up in its egonet.

Figure 41 - HiClus of Geo Distances; NOF

This figure refers to the Hierarchical Clustering of Geodesic Distances. The geodesic distance is the length of the shortest path between nodes. A hierarchical clustering of distances produces a tree like diagram which helps to understand which nodes are most similar to one another and it can easily be portrayed for a small number of nodes. However because of the comparatively large number of nodes the tree diagram here becomes confusing as the program has identified 98 clusters of varying sizes. The largest is that associated with NOF and this is shown in the figure. At first glance it seems similar to the NOF egonet but there are differences. Either way it shows that the NOF and its connections are important in a network sense.

Figure 42 - HiClus of Geo Distances; RenewTeesValley

A cluster associated with renewteesvalley was picked from the list of 98. As can be seen, by contrast with NOF it is rather sparse which again is interesting as this support organisation scored highly on degree centrality in Table 10 and particularly on a measure of out-degree centrality as it put a lot of links to third parties on its website. It would seem that the large proportion of isolates generated (as seen in the top left hand side of the figure) is due to many of its references having no other connections which depressed the magnitude of this cluster.

Calculated metrics and comment.

The following metrics are with reference to all 282 nodes and their associated linkages.

Table 14 shows the top 50 firms and other organisations sorted by highest eigenvector. The eigenvector approach is an effort to find the most central actors being those with the smallest ‘farness’ from others but in terms of the overall structure of the network. Betweenness centrality views an actor as being in a favoured position to the extent that the actor falls on the geodesic paths between the other pairs of actors in the network.

The use of closeness centrality emphasises the distance of an actor to all the others by focussing on the geodesic distance from each actor to all the others in the network. There is also column for centrality[34]. Using any of these metrics in the table the NOF is shown as having the greatest influence in the network both as a broker and as an influencer. This may be self evident from all the graphs but it is important to understand both the roles that it plays and the magnitude of such influence.

In addition to the above the following metrics were calculated for the network as a whole.

• Density, matrix average	0.0048
• Overall graph clustering coefficient	0.059
• Weighted overall graph clustering coefficient	0.022
• Maximum external ties	26962
• Maximum internal ties	52280
• E-I index	-0.033
• Average geodesic distance	4.306
• Distance based cohesion (compactness)	0.040

Looking first at the density and clustering coefficients. Compared to a typical social network with groups of friends these are very low figures at 0.0048 and just under 0.059 (6%) respectively. In an ideal world we would compare this with the whole of industry in the region to see if that 6% was greater than the regional average. Whilst 6% seems low it might in fact be high or even very high compared with a region largely unconnected internally. This would be a network form of location quotient. Unfortunately we live in the real world and at the present time it is not possible to have such a comparator. There are other industrial networks that have been studied [Kenny and Patton (2004)] with similar measures but unless these are in precisely the same sector or supposed cluster then such comparisons would not be valid. This is a theme that is returned to in Chapter 13.

[34] The centrality shown in the table is different from a simple link count as shown in earlier measures in Table 11. This because UCINET uses a more sophisticated measure which calculates the degree and normalized degree centrality of each vertex and gives the overall network degree centralization.

Table 14 - Multiple Measures of Centrality

ID	Organisation	Type	Degree	closeness	betweenness	eigenvector
19	nof	2	17.082	2.375	39.425	80.405
72	ata	1	6.05	2.354	6.653	32.173
153	mcnultyoffshore	1	3.915	2.353	3.45	28.483
54	ap-group	1	4.27	2.352	4.425	25.06
60	argonautics	2	4.982	2.348	4.799	24.721
246	setech-uk	1	2.491	2.351	3.546	22.566
2	Rigzone	3	6.05	2.336	4.788	22.554
275	wellstream	1	3.203	2.349	3.913	20.425
128	hos	1	2.847	2.353	4.635	19.282
115	fhpltd	1	3.559	2.347	5.335	18.959
70	ndi	2	2.847	2.33	1.111	17.663
103	engb	1	2.491	2.348	2.802	17.095
247	shepherdoffshore	1	3.203	2.339	3.447	16.866
279	wiltonmarine	1	2.135	2.343	2.02	16.695
9	ableuk	1	5.694	2.347	7.631	16.446
107	express-engineering	1	2.847	2.346	3.081	16.038
81	contractdesign	1	2.135	2.333	1.874	15.351
262	swintontechnology	1	2.491	2.345	3.225	15.218
46	amec	1	2.491	2.338	6.831	14.742
191	oceantecs	1	2.491	2.34	1.749	14.359
166	mkw	1	3.559	2.346	4.857	14.343
121	hedley-purvis	1	2.135	2.34	2.245	14.125
53	anson	1	1.068	2.335	0.219	14.037
172	nautronix	1	1.068	2.333	0.333	13.231
222	rbpipetech	1	1.068	2.333	0.387	13.115

269	tts-ltd	1	1.068	2.333	0.446	12.675
160	mjrcontrols	1	2.847	2.338	3.135	12.356
208	phoenixbeattie	1	1.423	2.332	0.581	12.178
148	marineprojectsint	1	1.423	2.339	1.359	11.809
202	penspenintegrity	1	2.135	2.33	2.425	11.75
44	akerkvaerner	1	2.135	2.323	1.825	11.741
95	dpt	1	2.491	2.337	3.156	11.682
48	anglitemp	1	2.135	2.333	2.203	11.55
223	rbvalvetech	1	1.068	2.337	0.58	10.92
157	metal-spinners	1	0.712	2.335	0.131	10.846
183	newarc	1	1.068	2.337	1.131	10.643
211	pipelineengineering	1	0.712	2.328	0.146	10.532
218	procladgroup	1	0.712	2.328	0.146	10.532
7	durham univ.	2	2.491	2.331	3.48	10.247
213	powerwholesale	1	1.423	2.331	2.699	10.189
220	qaweldtech	1	0.712	2.334	0.221	10.124
242	scientiasolutions	1	1.423	2.329	1.823	10.052
126	hiretorque	1	0.712	2.328	0.61	9.76
47	and-group	1	0.356	2.328	0	9.62
79	barriergroup	1	0.356	2.328	0	9.62
105	epigging	1	0.356	2.328	0	9.62
142	inductionbending	1	0.356	2.328	0	9.62
151	maxiflo	1	0.356	2.328	0	9.62
152	mckmid	1	0.356	2.328	0	9.62

The E-I index was looked at next as this is a useful measure of the group embedding based on comparing the numbers of ties within groups and between groups. For an index where all ties are within the group the maximum figure would be -1 and for all ties external to the group the index would be +1. Here we have a calculated E-I index of -0.033, a near balance of internal and external ties although under a random distribution the E-I index would be expected to have an index of -0.320 and hence a preponderance of internal ties.

The average geodesic distance (among reachable pairs) is 4.306, in other words on average any actor is just over four links from any other and that information may travel quickly through the network. In network terms this would seem to indicate a 'small world' network although for a full regional network, if it were possible to draw one this might not be the case. However the 'compactness' measure is low at 0.040. Although the graphs by observation appear to look like they might have dense networks with strong clustering the figures indicate otherwise and this is perhaps because we are dealing with industrial rather than people networks. In the latter the expectation is (and most examples show) that people networks have such metrics up to an order of magnitude larger than the figures found here.

As a final note on this sub section it is quite possible and indeed the temptation is there, to introduce a wide variety of metrics exploring the nature of both the overall network and the many subnets within it. However the pursuit of such further metrics has been stopped here on the basis that it is all too easy to become detached from the objective of answering the research question. The graphs and metrics discerned to this point are sufficient to gain an understanding of what has been done and what could be done in the context of supporting internet based methods of finding industrial networks and clusters and actually measuring them using metrics from social networks analysis.

Actual comparison with similar industrial cluster structures is of course another matter and this possibility is discussed in the final conclusions.

11.3 Geographic Location

The research question is relatively open in terms of its scope and application and indeed refers to the internet as an ‘information resource’. Thus far the thesis has concentrated on using the internet to find and use data in the form of text and also for link derived information which is the subject of this chapter. There are however other possibilities for contributing to the study of for example agglomerated industrial activity and one such area is the use of geo-referencing tools for use in applications where location is important.

In most cluster studies undertaken up to at least the last few years the location of firms was deemed important or even a pre requisite. Although work on virtual clusters and the use of the internet for distance independent collaboration has been looked at particularly for digital industries the issue of geography is a crucial parameter particularly for industries and sectors that may have some historical connection with an area. As we are looking here at a region which by definition is bounded geographically the actual location of firms that are in a network of common interest may not be so visible on the basis that every firm is relatively near to all the others anyway. Nevertheless it was thought of some interest to compare the Euclidian form of networks looked at so far with the actual geographical location of the same members. To undertake this we have used as the basis the combined Subsea database is composed of some 282 members including first level contacts that are part of the network as shown from Figure 30 onwards.

Method of Approach

There are a number of ways of geo-coding company addresses and portraying the actual location of these on a suitable map. The usual way is by converting a 7 figure postcode into a Lat/Long co-ordinate and placing a suitable graphic at the location point. This is a relatively simple method and works well enough except in two circumstances. Firstly a postcode conversion takes as its location the postcode area centroid and for a rural setting the postcode area around the centroid might be rather large with the firm location anywhere within that area. It may be therefore some distance from the actual centroid. Secondly for business parks with multiple businesses they might all have the same postcode and hence the same plot location.

Dense concentrations of firms either do not show up or special arrangements have to be made to emphasis the fact that there are many firms in the same location.

The method used here was slightly different and again internet derived tools now available were used. The basic method used was as follows:

1. Use URL from database to look up the most appropriate regional address.
2. Using the full postal address use Google Earth™ to locate the actual geographic space occupied by the company or organisation. This method, with a little intelligence on the part of the observer can cope with office blocks, business centres, business parks and like where there are a number of businesses with ostensibly the same postcode or even address.
3. Use the locational readout from Google Earth to give Lat/Long readings of the found address.
4. Reassemble database of URLs with company name and Lat/long co-ordinates.
5. Use this database as input into a suitable mapping programme. In this case Microsoft MapPoint was used as in addition to being able to accept most common database formats for a file of locations with additional data fields it is easy to accurately portray topographical features such as rivers, roads, airports, boundaries and other urban features which might have a bearing on the historical location of certain types of firm e.g. Marine.

The net result of this exercise is as shown in Figure 43 below. As the physical dimensions of the map are 100km by 120km covering an area of 1200km2 the detail is not evident although particular loci of activity can be seen around the Tyne and Wear areas together with a lesser one centred on the river Tees. By expanding the Tyne and Wear area to twice the size it is quite possible to discern the possible influence of the various topographical features of the landscape and also others such as the presence of Universities and the airport. This is shown in Figure 44. It might be argued that with such a small region whatever sector was looked most industrial activity is going to be based somewhere in the urban areas and that for Subsea might look very similar to the layout for say precision engineering. One of the benefits of modern mapping systems is that precise locations and the surrounding attributes of the firm's location can be examined. For some firms in the Subsea segment historical antecedents in the river based industries may have played a part in location but those engaged in design often have a pragmatic preference to be 'somewhere near the action' and this can also be seen with groups of marine design contributors not far removed geographically from the constructors of Subsea artefacts with many of the latter actually needing access to the waterfront. Although the Figures shown here are static, in the program used they can be dynamic with each location having within it embedded attributes ranging from database information to aerial photographs and of course the links established in earlier work. As an exercise some of these links were plotted on the maps to show the location of firms[35] and their network connections. However it very quickly became obvious that the resulting mesh combined with all the map attributes resulted a very unwieldy diagram in which it was almost impossible to discern either company location or network relationship. This form of diagram was therefore not continued.

[35] Many of the firms used in the Subsea examination, whilst having a regional contribution or connection did not in fact actually have a regional address. Examples are the industry websites such as www.rigzone.com or other temporary organisations such as major rig projects designated by the destination oilfield. The actual number of firms used therefore was less than the 283 on the original network database.

Figure 43 - Location of Subsea Activity, North East Region

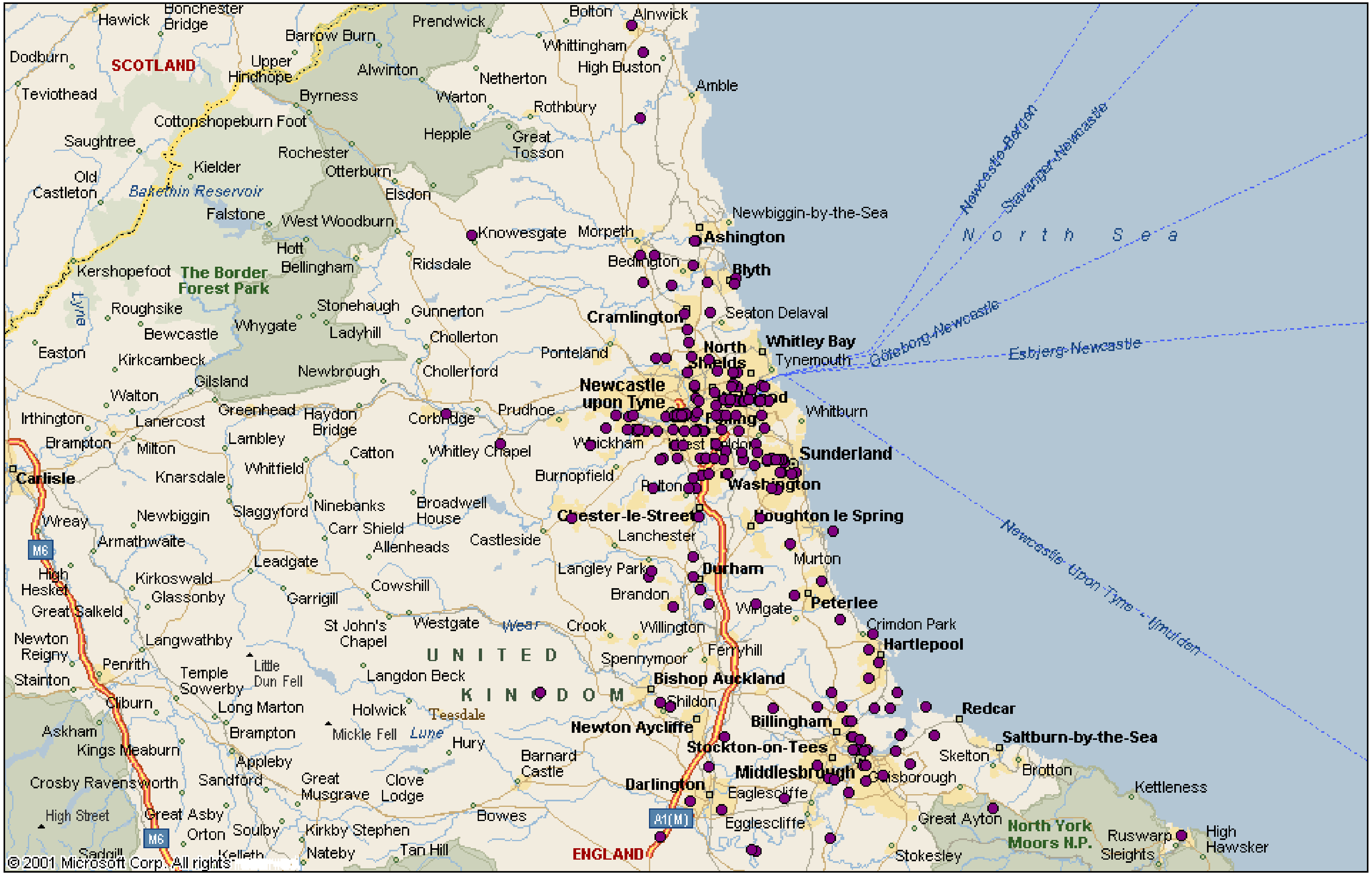
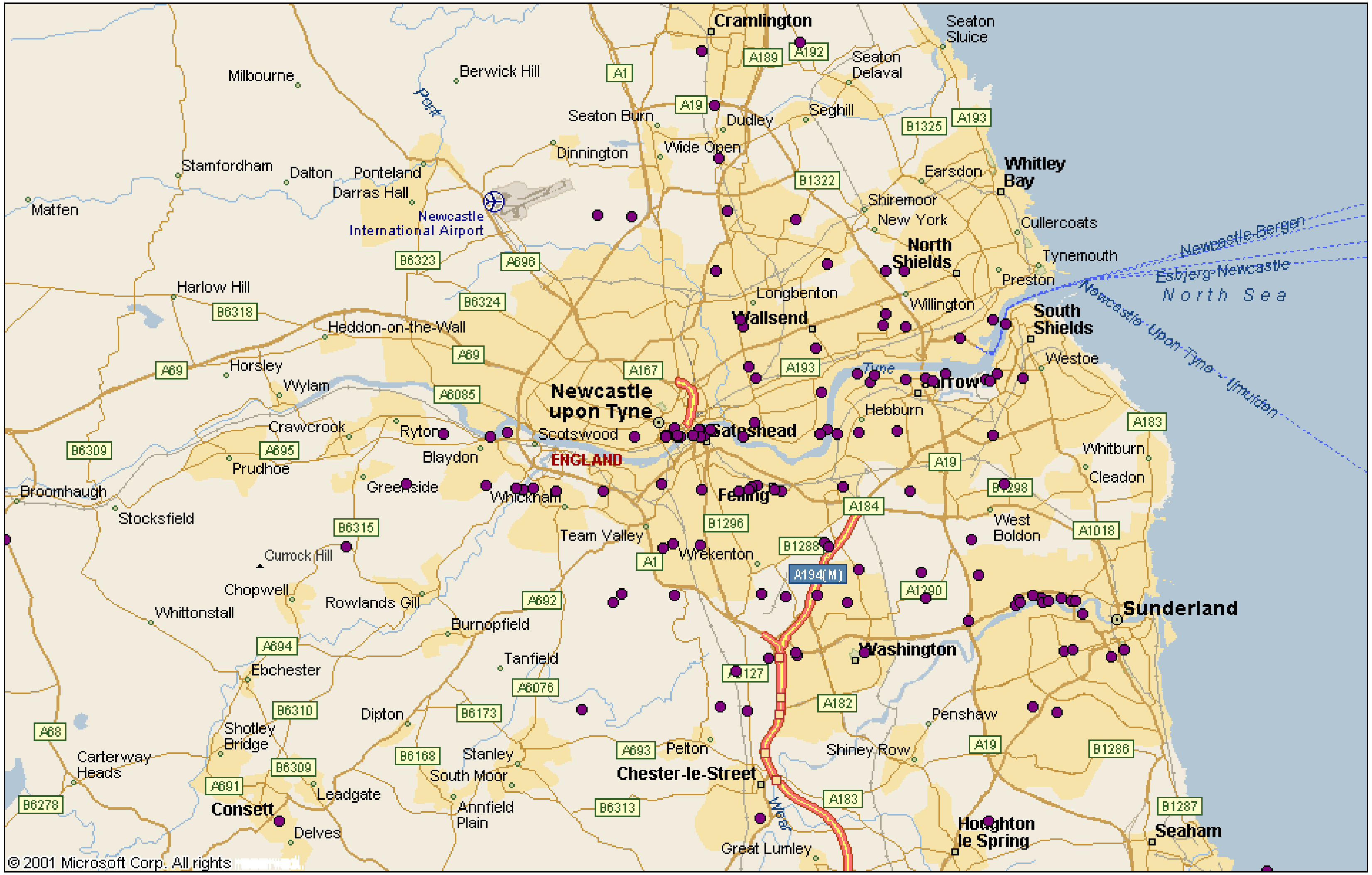


Figure 44 - Subsea related organisations located in Tyne and Wear



11.4 Conclusion of work on web derived linkages

As would be expected this chapter has occupied a significant part of the thesis. After consideration of networks generally which was looked at in a literature review in Chapter 10 this chapter started with a preliminary look at network thinking and how this might apply to the research question. In the same way that the chapters on finding text strings by using the internet applied to a large corpus of firms looked at existing tools so this section similarly looked for methods that could be used or modified to make progress towards the goal of finding interconnections between firms. The early part of this research was therefore concerned with an investigation into available tools that helped to discern linkages between firms. With link finding programs however it became clear that if we were to look at links between firms and further if such firms were wholly within a defined geographical region there did not appear to be any such programs available either commercially or in the academic sphere. This is not surprising as what was being sought was highly specialised and not the sort of development that any commercial firm would invest in bringing to market. On the basis of this it was then decided to design and implement a custom spider based program to search automatically for the presence of hyperlinks on any site that linked to any site or sites within a named population of URLs. This whole exercise and in particular the 'tuning' of the program to deal with the many variants of internal website design and programming absorbed a significant number of man-hours and indeed at one point the programming difficulties seemed so problematical that the whole notion of finding a good sample of representative links within a geographically bounded set of URLs was in some doubt. Continued testing and development removed these problems to an acceptable degree where the term 'acceptable' in this case could be defined as a compromise between continued and ultimately lengthy further work and the solving of a reducing number of remaining problem areas. With the program finally functioning however it did fulfil its objective of finding links but with some restrictions. The restrictions were primarily those associated with the size of the corpus of firms under consideration.

Before any real testing had been carried out with the link finding spider it was initially thought that the entire available corpus of URLs (over 10000) at the time might be spidered to 'see what emerged' in the way of connected firms. This would be analogous to social activity amongst groups of friends but which here relate to some sort of business connection between firms. There are two drawbacks with this approach that were confirmed with testing. Firstly as with the groups of friends the links observed may be due to the most socially active or in the case of firms those who choose to put links to other organisations of interest on their company websites. Secondly the sheer size of the data generated when spidering links to a specified depth and the manipulation of the resultant data. Obviously even a simple 2D matrix with 100 URLs generates a matrix of $1*10^4$ cells and for the full and final URL database acquired it would be $(14*10^3)^2 = 1.96*10^7$ cells, a problem touched on in Section 10.6.2. As a result of testing with a discrete but increasing number of URL cohorts it was possible to extrapolate to gain a view on what might be a practical limit for in-region link spidering and it was shown that for cohorts much greater than around 500 in number the time to spider would be significant. It was at this point that the idea of concentrating on groups of firms that were less than 500 in number as derived from the RBKS was taken up. In the literature review Communities of Practice had been looked at briefly on the basis that an understanding of their origins and functionality would help the research. Although there are a number of definitions of Cs of P in the same way that different authors have a view as to what constitutes a cluster it seemed that a C of P based on one or more professions might be a good candidate grouping to look for overt associations on the basis that professions are like a club where many people know each other socially as well as through trading relationships and this social aspect might be translated into website hyperlinks.

As has been noted in comment on the results, in the event the outcome of spidering even quite a large group who had words commencing with or containing the word ‘architect’ yielded very little in the way of embedded hyperlinks and the possible reasons for this were discussed. In the methodology it had always been the intention to look at any way possible by using the power of the internet to gain information on the firm, the firm’s activities and the general industrial milieu in which the firm operates and so the task of examining firm websites for other clues to associations was undertaken. This further work on examining company websites for visual clues to a firm’s associations turned out to be a welcome addition even though it was a time consuming exercise rather than a particularly complex one.

The next part of the investigation involved the use of in-links and again the further addition of this type of link transformed the node and link population for the main cohort used as the demonstrator being that associated with subsea activities. Clearly the addition of embedded links, observed links and in-links gives a much richer picture of the associations between common interest firms. The additional opportunity to compare the connections in a cluster undertaken by a third party using snowball sampling and associated interviews with connections derived entirely by web based means is valuable even though the match achieved is, at a maximum of the order of 50% between the two methods.

It is tempting to become seduced by the very nature of the network diagrams and the interesting looking connections between quite a large number of nodes. It is however important to keep a sense of perspective as many of the firms found as feedstock for input to link finding, either by program or manually and on which the whole of the data on which the network maps, comment and conclusions are subsequently based suffers from general internet quality and bias problems which have been discussed in the context of spidering for text in preceding chapters although perhaps not to the same extent. The quality problems with linkages are subtly different in that with text there is the whole problem of noise emanating from inappropriate sites together with other texts which may be in context but of doubtful veracity. In the manner of the quality of linkages from one site to another the first problem, that of noise is largely absent and we are left with the issue of whether the link is ‘correct’ for the site being looked at. The link was of course decided upon by a person but again there is unlikely to be a disinterested party checking such links as to accuracy, appropriateness or indeed truthfulness. Thus a person intent on portraying their website as having links to all manner of ‘interesting’ other sites could probably do so unless and until some third party objected. With a little observation of website structures it would appear that this sort of thing is quite common for small and often new firms who push the literal boundaries to give the impression that their networks are rather larger than they actually are. The same phenomena do not seem to occur with larger corporate bodies who presumably have a full time PR department and have checks and balances in place to prevent this happening. They are also usually at a stage in their development where they have enough company and client information and experience to have a website composed entirely of factual information.

At a late stage in the thesis, in spite of some misgivings about the use of a powerful metrics calculation program it was decided to submit the data for the subsea network to UCINET although the metrics determined were limited to those derived from link count and direction only. The result of these analyses did in fact yield useful information particularly in determining some significant nodes by metrics other than by simple link count based centrality. In addition it was possible to determine mathematical descriptions of the various descriptors of the network such as clustering co-efficients and measure of connectivity between nodes. This did however highlight a problem in that comparisons with other like-for-like networks were difficult to find, the nearest examples seeming to be either for people type networks [Wassermann and Faust (1994)] or the internet [Chakrabarti *et al*. (1999)].

The penultimate part of this section on link finding involved a comparison of links determined manually by the possibly robust but very time consuming process of asking key respondents in the individual firms who they traded with either for artefacts or for knowledge and then comparing these results with a similar exercise done, not in person, but by internet derived linkages. One of the metrics used for comparison was degree centrality and for the sub-sea sector looked at the results were that c.50% of the links discerned manually could be found by internet means alone.

Again all the comments made in the context of finding firms through text spidering regarding the size of the sample of 14000 URLs apply in equal measure to link finding. Although it cannot presently be proved, if in the future it were possible to obtain a full set of URLs for the region and firms put more links on their sites in the manner of www.sevenrings.co.uk[36] then it is certain that both the number of nodes and the number of links would rise for any kind of grouping whether it was subsequently found to be a significant cluster or otherwise.

As with web mining the technology for link analysis is still evolving and it is likely that tools and other methods at present unknown will be developed to support this area of investigation.

A further observation that was noted for the subsea segment was that very large organisations such as AMEC and Shell had few links on their site as a proportion of their activities which cover a huge range of project in many parts of the world and

[36] Page 188

for some projects the country scarcely gets a mention let alone a small English region. Thus all the local subcontractors who may have worked on a significant AMEC or Shell brokered project are not found in any way through the websites of these organisations. Conversely however many of those same subcontractors do reference large projects they have worked on even though, in the scheme of things, their contribution may have been modest. Thus the link finding system would seem to favour the identification of smaller firms, a point that is discussed further in the overall conclusions.

The final part of this chapter has been concerned with plotting the actual physical location of the regional subsea sector firms. In some ways this is a more traditional way of observing the possibility of clustering on a geographic basis in that it is possible to see by observation the effects of proximity. However the method of plotting used here is also a demonstration of other new internet derived tools, the primary means of enablement being Google Earth. This program was first made available by Google in rudimentary form in 2004 but has been developed somewhat since that time[37]. The purpose of this exercise was twofold. Firstly to undertake a conventional locational plot of the location of the segment determined by web based means and secondly as a demonstrator as to how a high degree of precision in the firm's location can be obtained by using a publicly available mapping tool as a geocoding system. The conclusion is that such tools work well although some manual input is required to adjust address location to actual location for companies that are very close together. As has been noted above what did not work so well was a comparison of the derived mesh of Figure 32 when translated into geographic links for the simple reason that the presence of so many linking lines in a relatively small

[37] http://www.google.com/corporate/history.html

Google Earth has only recently been available in high resolution format for significant parts of the earth's surface including the North East of England. Apart from the ability to use it to establish very precise locations it is also perhaps of interest for visibly exploring those locations that are a domicile for certain types of firm. Thus whilst for example head office based administrative activities can take place anywhere, for an offshore construction over a certain size ready access to a deep water terminal is required. Whilst this may seem an obvious application it helps to build a picture and other less obvious applications might be internet based retailers being located with a certain travelling time of an airport or a UPS depot plot area largely obscured the nodes, particularly for example along the banks of the river Tyne.

This problem would not necessarily be an issue for a larger area with fewer nodes and less clustering.

Chapter 12. Investigations into Overlaps Between Clusters

12.1 Preamble

In the preceding investigations of Chapter 9 it was observed that many of the firms identified by the RBKS for a single grouping such as 'Defence' appeared in other groups such as 'Offshore'. At the time this phenomenon was merely noted as other analyses were being carried out but this chapter looks at the degree to which these firms might appear in more than one list or group membership. This might be important for two reasons;

1. Firstly one of the conclusions from the literature review was that many firms are engaged in more than one market and that some of the shortcomings associated with the use of Primary and Secondary SICs resulted in the full spectrum of a firm's activities not being fully captured. The RBKS is of course indifferent to SICs as it picks up all activities on a spidered site. By looking for firms which appear on a number of 'cluster' lists, by market or by area of activity or competence would be a useful addition to knowledge or methodology and for, in our case a profile of the N.E. region.
2. Secondly in the matter of commonality of networks. In the same way that a firm undertakes different types of activity or a similar but modified activity in different markets is the firm a member of different networks for different markets or different technologies, services or manufacturing processes?

In view of these aspects a short test was designed to examine the degree to which such assumed commonalities occurred.

12.2 Design of the Test

In earlier work related to both Chapter 9 and Section 11.1 there was a requirement to find groups of firms that appeared from their websites to be engaged in some common activity. These covered a range of sectors as, particularly with link finding, evidence of collaborations were being sought, particularly from those firms that might be regarded as being part of a community of practice. As has already been noted in Section 10.6.2, this was not particularly successful and firms that were expected to show some evidence of networking amongst the C of P (if it existed) did not do so, at least not using the presence of regional embedded links as the indicator. However for some of these groups that were examined and in particular those in

- Marine (199)
- Environmental (386)
- Offshore (344)
- Subsea (283)
- Defence (290)

it was noticed that many of the same company names kept appearing in one or more sub sectors. As the data for these cohorts were available as single field indexes it was decided to merge these indexes suitably keyed to discern which appeared in more than one list. The number of firms in each of the cohorts above is shown in brackets.

The method of approach was to assign each of the five cohorts a unique colour and then to merge and sort all the lists by company name. By observation it could then be seen which firms appeared in more than one list. This gave a total population of 1502 but of course this figure included a number of duplications, the exact figure being determined by examination later in this chapter.

Results

The basic results of this exercise are shown below in Table 15.

Line #1 shows the population used for each of the segments being Marine, Environmental, Offshore, Subsea and Defence totalling some 1502 firms. All were derived from the RKBS with exception that 'Subsea' was augmented with additional firms as described in Section 11.1.5. Of these 1502 firms 198 of them had activity in other sectors other than their own (#2 and line #5).

Line #4 shows that of these firms active in other sectors 140 were active in one other segment, 47 had activity in 2 segments, 10 were in three other segments and one firm was active in an additional 4 segments.

Line #3 shows the proportion, by sector, of activity in other segments. The highest is 'Defence' at 42% and the lowest is 'Subsea' at 4 %. The average across all five segments is 18% of firms in one sector being engaged in one or more other segments.

Lines 7-11 inclusive show the metrics by each individual segment to segment in terms of the number of firms thus engaged.

Table 15 - Degree of Involvement by Firms in other Sectors

		Main Sector						**Count in other sectors**			
#		**Marine**	**Env**	**Offshore**	**Subsea**	**Defence**	**Totals**	**+1**	**+2**	**+3**	**+4**
1	Populations	199	386	344	283	290	1502				
2	Firms in other sectors	49	45	40	11	123	268				
3	% 'hits'	25%	12%	12%	4%	42%	18%				
4	No. of firms in additional sectors							140	47	10	1
5	Cumulative No. of firms							140	187	197	198
6	Sector by sector detail:										
7	Marine		22	4	1	30					
8	Env	1		1	3	29					
9	Offshore	29	15		6	38					
10	Subsea	18	8	31		26					
11	Defence	1	0	4	1						
12	**Totals**	**49**	**45**	**40**	**11**	**123**	**268**				
	Total number of firms cited outside base sector		**268**								

12.3 Commentary and Conclusions

Marine – For firms in the Marine sector who have an involvement in one or more of the five segments not surprisingly the majority of these are active in the Offshore segment with a significant proportion citing Subsea as one of their areas of activity. They appear not to have moved into Environmental activities as yet but perhaps most surprisingly only one shows Defence as a market or area of interest. However the 'direction' of interest may be important as when we look at Defence as a sector there is considerable involvement in Marine.

Environmental – This sector covers the whole gamut from Environmental Engineering through to lawyers involved in asbestos related litigation. It was noted however on page 187 that most of the activities found using the word 'Environmental' which were felt unlikely to add to the industrial milieu such as those found on the Corporate Social Responsibility Statements of many firms had been removed. At 12% engagement those in the environmental industries have some activity in other areas the most being in Marine and Offshore, both segments which are coming under increasing pressure from environmental legislation.

Offshore – There are few surprises here in that firms in the Offshore sector have a majority interest in Subsea with relatively few in the others.

Subsea – For firms who are engaged in subsea the figures indicate a highly specialised activity with few firms who specialise in this segment engaging in the other sectors. At 4 % overall this is by far the lowest proportion of the five sectors looked at.

Defence – Firms in this sector show a significant involvement in the other four. This is an interesting result because when the early work on comparative methods of finding possible candidate clusters in the region was being examined very few firms showed up as being in the defence sector at all as determined by SIC count (see Table 4). Here we have a population of 290 firms engaged in some Defence related activity and of these some 42% of firms between them are engaged in 123 incidences in at least one of the other four segments.

A word of caution. The information determined by the RBKS does not, unlike an SIC based system, name a primary and other secondary activities and the internal ranking system of the RBKS is used to determine which is the most likely main line of business for the firm. This ranking system as discussed previously is a function of the number of times a chosen word appears on a website. The ranking feature is something of a two edged sword for the RBKS in that on the one hand it does find many activities that the firm is engaged in but it does not necessarily 'know' which is the most 'important'. It is entirely possible for a highly focussed firm in say subcontract parts for armaments to only mention the word 'defence' once or twice on its website. Conversely a company website designer specialising in the defence sector may have many references to the word 'defence' on its own website thereby achieving a higher ranking in the RBKS so it was felt that the ranking system internal to the RKBS and which was part of the original Phantom program was best ignored for the purposes of finding firms engaged in the various segments. In an SIC based system the originators would of course put down the company's main activity as a Primary SIC although if the firm had several activities all of equal importance this would be problematical. So in this case when looking at the involvement of a firm in different sectors we have, for example the case where www.powerconv.alstom.com is one that cites all five sectors as an area of interest and therefore presumably competence. The company is well known in the power generation and high power electrical propulsion business. Do we therefore say it is a Marine company with some secondary activity in the other segments or an Offshore technology supplier with involvement in the other sectors. Using the RBKS this type of SIC constrained thinking is not necessary and the company simply addresses those markets and processes competencies shown on its website. This is a theme that is taken up in the overall conclusions.

A further comment is that here we have looked at the cross fertilisation between 5 sectors or segments. However some of the firms will also be engaged in other segments and if say 'Mechanical Engineering' or 'Process Engineering' or 'Pharmaceuticals' or 'Energy and Power' had been included then it is likely that many of the population of firms used here would show up in those segments also.

When writing this section and particularly with respect to the comment made above for Alstom Power it was noted that:

1. The five segments used here do have some commonality of process. They mostly are involved directly or in a support function for engineering, often for larger structures, they require project management skills involving at least medium and sometimes high complexity engineering, their products operate in difficult or even hostile environments and they require facilities for fabrication, machining and assembly using or being part of first or second tier subcontractors. In addition to this commonality of process all the segments are subject to strict Quality Assurance protocols and increasing legislative pressure regarding Health and Safety and environmental protection as indicated by the presence in the network of various government and other agencies concerned with these aspects.
2. From the work done here it would appear that there are a group of firms in the region that have such skills and experience to service one or more of these five segments and even acknowledging that the RBKS is but a sample these firms would seem to number at least 268. For the entire region the true figure of course will be greater.
3. In many of the definitions of a cluster shown in Appendix 3. Clusters: A Variety of Definitions, although a few commentators refer to 'single industries' most acknowledge firms' interdependencies and localised networks of producers with a regional dimension. Here we seem to have a group of firms that do fit with these definitions except that there is also a concentration of a particular set of competencies as much as there is in physical 'clusters' specifically in Defence, Offshore, Environmental technologies, Subsea or Marine.

Chapter 13. Summary, Conclusions from the Research and Recommendations for Further Work

This chapter is organised on the basis of an initial short review of the thesis from literature search through to final network analyses. It then addresses, in the context of the research question three key issues of:

- What works
- What doesn't work
- What could work (given more research)

Existing research and associated literature

The literature review was driven by the methodology mapped out in the first chapter in that in order to understand how the internet might contribute to the enhancement of our understanding of industrial clusters it would be necessary to have some knowledge of:

1. The notion and use of industrial clusters, their historical background, supporting economic theory and the present state of knowledge including gaps in the literature and shortcomings in practice.
2. The use of the internet as a valid information resource
3. Networks as both a mathematical construct and for practical utility

The first part, that of the background to industrial cluster research elicited a considerable corpus of papers and associated literature as the subject has been of significant interest, mostly over the last twenty years, to both researchers and also practitioners charged with economic development on a regional or national stage. The driver for such study of such clusters in the widest industry sense has been to try and understand the rich networks of interaction between buyers and suppliers, between supporting organisations, the inferred untraded interdependencies and with the external environment. A key requirement to underpin such research and subsequent use has been the need for comprehensive lists of industrial firms and supporting organisations in the geographical area under study. Most of these lists have within their individual records, fields describing the activity of the firm usually through the medium of a national SIC scheme accompanied by a short text descriptor.

From the literature a number of frequently cited researchers have indicated a problem area where further study was required and this is related to the fact that many companies undertake more than one activity in more than a single market but that such diversity was difficult to discern using conventional SIC based company lists. In addition the SIC system itself had difficulty in accommodating companies specialising in new or developing technologies. The complaint was that because of these shortcomings *"significant clusters may be obscured or even go unrecognised"*.

The next part of the literature survey, again in line with the methodology of seeking knowledge and tools that would support subsequent investigation in the context of the research question was that of information extraction from the world wide web. Whilst it is self evident that the WWW is a hypertext corpus of immense size the problems of addressing such an unregulated and non-standardised medium was an issue that had been identified and addressed by a number of researchers if not entirely solved. It was also noted that much of both the research background and publicly available programs for extracting web based information had been developed largely for marketing purposes. The initial reviews looked firstly at the research background for extracting text or possibly information from the web and this in turn was followed by a critique of possible tools and other programs available for undertaking such a task. It was concluded that the research to date had concentrated on methodology but had only been applied in an industrial research context largely to demonstrate the worth of a particular program rather than as a useful adaptation that could help progress the research question here. As a result of these findings it was decided to develop a new program that could extract useful information on the firm using web based means. It was also acknowledged that new internet based search tools of various kinds were under always under development by third parties and as they appeared the most promising of these were examined with a view to adaptation concurrently with other work being carried out in this research.

The final main part of the literature survey was that concerned with networks. It was recognised in the literature on industrial clustering that a key element was the association between organisations, in other words the degree to which such associations could be regarded as a network. Network theory has a long history in the mathematical sciences but the availability of cheap computing power has reawakened research into applications for a wide variety of human and biological systems. The literature therefore was looked at both in the classical sense and also with research into applications which included the internet which is of course a network albeit one of enormous size. The literature gave very wide ranging examples but it was possible to narrow these down to the key elements of scale free or small world networks together with the place of hubs and authorities that would be appropriate for the modelling of industrial linkages.

The key messages from the literature search, both in terms of building a methodological framework and of identifying gaps in the literature were;

- The term ‘industrial cluster’ although widely defined in the literature is still subject to debate particularly as regards economic worth.
- The use of SICs to describe the firm had some negative implications for cluster research in that significant sections of industry and commerce could go unrecognised.
- It was not easy to identify untraded interdependencies and in any case the definition of these varies according to different authorities.
- The internet, whilst being an information resource of immense size is not ideally suited for automated information extraction because of way individual websites are constructed and the presence of large amounts of irrelevant ‘noise’.
- Information finding programs such as Google Local are difficult to control in a research sense.
- Web link structures have been subject to a significant research effort and although interesting from a pure research viewpoint few have been concerned with application to industry other than as support for marketing tools.

<u>The Research Process</u>

The actual research i.e. that which is new began with Chapter 7 and based upon the conclusions from the literature survey a number of tools were developed to try and gain information about potential cluster populations using information derived entirely from the web. The objectives were split broadly into two parts the first being the finding of information from company or other organisational websites in order to gain a comprehensive view of the activities of the firm or organisation. The second part was to use the internet to look at the ties that form networks of trading and knowledge relationship between companies in any supposed industrial cluster. A key underpinning element of this part of the research was the establishment of a representative cohort of regional company URLs that would ideally be large enough to give some degree of confidence in the methodologies subsequently developed. In the event the number of unique validated URLs acquired reached some 14000 with the process of acquiring such URLs by a variety of means, having continued throughout the period of the research.

What has worked

This section looks at those aspects of the research which have worked well and have clearly contributed both to knowledge and practice. In order of their undertaking they are seen as:

1. Keyword searching from a large number of spidered websites - There is a degree of confidence in this process because it picks up the combined information base of as many websites as have been spidered. The result is much richer than anything that can be obtained from SIC based searching and this has been demonstrated by comparison with data of known provenance from a credit reference database. The process is particularly good for specialised activities or for those hard to find industrial functions which may have no overt SIC. The results can be used in a variety of ways including the augmenting of a conventional industrial database.
2. Some indication of networks of activity – the discerning of networks in an industrial context is not perfect and should not be given undue credence but the connections that have been found between firms do give some indication of the milieu associated with any particular group in a defined location.
3. An ability to find evidence of clustering according to one or more of the commonly used definitions by a combination of keyword search, link discovery and manual methods.
4. An ability to find evidence of other types of commonalities such as process or competence.

What did not work

In some respects the elements of the research that did not yield the expected contribution are not wholly on the critical path to answering the research question or adding to the body of knowledge. The two most obvious examples are the attempts to get the URL Scraper and the Website finder to function efficiently. Both of these programs occupied significant amounts of both the author’s time and also required external advice from contributors in the field. In the case of the URL Scraper program although it was initially seen as a key element for obtaining information from ‘scraped’ websites the general research policy of following several concurrent lines of investigation in the event yielded better results from a modified commercial program this being ‘Phantom’. The lesson to be learned from the exercise however was that text mining from the internet by whatever means is a non-trivial exercise and due arrangements have to be incorporated to deal with the fact that the internet is non-standardised, unregulated and has millions of authors each with their own ideas of how a website should be constructed.

The Website or URL finder program similarly was an ‘off-line’ attempt to speed up the process of acquiring larger numbers of regional URLs from available company lists. Here the difficulties encountered were caused by the presence of a number of catalogue sites or integrators being the entry point for company information rather than the company URL itself. This problem could probably have been solved eventually but it was clear that to continue would not have been a sensible use of time. On a number of occasions when dealing with internet related problems it had been important to remember that this thesis is about answering the research question, not about getting deeply involved in solving programming issues, particularly those that could be circumvented by other means.

What could work better.

It has been noted many times in the main body of the work that the URL database used is but a sample of industry and commerce in the region but it was felt to be sufficiently large enough to demonstrate the worth of the various information acquisition and network schemes that were tried out. Because the number of URLs eventually became quite significant (c.14000) there has always been a temptation to try out various forms of search and to look at different network schema, a position that may not be entirely warranted given the remarks in the body of the text regarding biases. Clearly if we had perfect knowledge of 'what goes on' in industry in a locale by the simple expedient of having all the URLs available then such a position would be justified. An obvious way to acquire these, in very practical terms would be a 'sledgehammer' approach whereby a contact centre of say 20 people could be used to scan every directory, yellow pages, company list, companies house register available together with the specialised use of Google Local as described on page 139 followed by the use of a search engine to look up each company URL. At a rough estimate it should be possible to get through 3000-5000 URLs per day dependent on the number of operatives involved. For a region the size of the North East of England therefore if company lists were available or could be acquired for the majority of firms in the region the majority of URLs could be thus captured in approximately 2 weeks.

The other area that perhaps needs further research, and this is noted in recommendations for further research starting on page 292, is a better understanding of link relationships. At present no particular form of relationship between one web site and another has been assumed other than directionality derived from whether an in-link or out-link is present. It is therefore not always known for certain if a link relates to a buyer, a supplier or some sort of knowledge exchange or a simple information link although in many instances the form of relationship may be guessed at. Because of this limitation the links cannot be ascribed properties and the 'strength' of the link is not determined. Consequently any interesting network metrics that require the use of a measure of link strength cannot be used. It may be possible to have some rudimentary ranking system for each link such as a key link, average weight, peripheral (non critical), information only or unknown.

Finally there are the general aspects associated with the process of finding text and links and the sorting of the resultant database as in the RBKS. Text spidering using 'Phantom' takes a long time and a faster spider using parallel search processes would be required for say whole country spidering. The algorithms driving Phantom and hence the RBKS are adequate but there may be other faster ways of spidering material and concurrently removing noise and irrelevant text and other characters on the fly. Additionally when searching for keywords or phrases the Boolean search facility in Phantom is relatively simplistic and a Google type search engine or a modern natural language engine such as Powerset[38] or TrueKnowledge[39] would be an advance enabling more precise searching of the spidered material database. Since the original work was done using Phantom and its predecessors a number of products have started to appear on the market and which claim to have a similar capability with some additional functions to control the search and subsequent output.

Critical Appraisal

In carrying out a critical appraisal of the research there are a small number of key issues that bear upon the final outcomes together with a number of assumptions and approximations that have had to be made as the research progressed.

The first fundamental issue is that related to the way that both the text for the RBKS and the acquisition of outgoing linkages are wholly dependent upon what a company puts on its website. Both of these methods acquire information from companies that 'promote' themselves through their website. To promote in this context is to show your clients, collaborators, suppliers and other related and supporting linkages on your website or indeed to have lots of links generally of various kinds and also if possible to have others link back to your site. This process however is not always related to the economic contribution a company might make or the key role that a broker of information might perform. Generally speaking when searching for text strings on a large corpus of spidered sites this was not so much of a problem although some large firms (e.g. www.pwcglobal.com) had to be excluded from the process because of the very large number of pages being acquired and indexed by the spiders. The variation between spidered websites therefore could be very 'uneven' in terms of spidered material and this was not necessarily correlated with size of firm – more with size of website.

The second fundamental problem was related to the observation of links between organisations in that again the number of out-links on any web site was dependent on what the website author chose to put on the site. Large numbers of company websites have no out-links at all. Although this may change in the future it again shows up the unevenness of observed links with the actual presumed collaborative activity (or lack of it) for the organisation. This is a drawback that has been known

[38] http://www.powerset.com/about

[39] http://www.trueknowledge.com/

about and indeed was discussed particularly with respect to the websites of very large corporates when looking at the subsea sector. Here it was quite possible for the links from a relatively small subcontractor to exceed in number those of a large and key company. In-links however are somewhat different in that a third party not involved with the website design of the subject firm decides to reference that firm on their own website but again there will be a degree of unevenness of the referencing process due to individual authorship of the referencing website.

Other sources of inaccuracies, as discussed could be website authors not being entirely truthful regarding their firm’s capabilities and interests on the basis that, as long as any third parties are not actually defamed, the consequences for the firm (probably small firms) are likely to be minor.

The problem of definition of the term ‘untraded interdependencies’ was discussed early on in Chapter 2. When using the internet in the way shown in the thesis to find companies or other organisations or groups of these it is much easier to achieve this in terms of Porter’s ‘related and supporting industries’ that it is in terms of a strict Storper based definition of untraded interdependencies. In examining the network therefore we have looked at the part played by some of these non-traded entities or at least those not operating on a fully commercial basis. Examples that appear prominently are the trade associations, government departments, the educational sector, business support agencies, not-for-profit research centres and various regional associations charged with helping industry to grow and prosper. Whilst these organisations are undoubtedly supporting industries it is arguable whether they can be regarded as a proxy for the benefits of agglomeration which are not the result of direct trading relationships between firms and it is thus still an open question as to whether true untraded interdependencies have been positively identified.

Has the research question been answered?

The research question is:

‘Can the use of the internet and the world wide web as an information resource add anything useful to more conventional methods of researching industrial clusters and networks?’

In some respects the question was deliberately framed somewhat loosely as at the outset it was difficult to predict the level of contribution that the internet would make but it would be surprising if use of the internet and the web could not contribute at least something to industrial cluster and network research. However the assumption throughout has been that this contribution should be either unique or substantive or indeed both. The section above on What has worked does summarise the main benefits that have accrued from the research in the context of the research question. The use of internet based searching for company data and information it would appear is indeed a rich source of gaining a comprehensive view of the activity of organisations in a defined geographical area provided of course that a substantive population of seed URLs can be obtained in the first place. The resultant information database extracted from text harvesting from these URLs can be used to search for similarities not just in types of product or service but also for common areas of competence or of process. In addition supporting organisations can be found that are, in many cases, characteristic of the attributes expected of an industrial cluster. It was always expected and indeed is implied in the research question that the information found using the internet would be an adjunct rather than a replacement for conventional cluster studies and this has been borne out with the comparative tests undertaken. Whilst the RKBS provides a rich vocabulary of descriptive text it says nothing directly about the size of firms, the financials or the employment. However as these kind of data are available from company lists such as those provided by DNB or Experian[40] the two methods of search for company data can be highly complementary although it has been pointed out that the RKBS does also find

[40] http://www.experiangroup.com/ Experian “is a global information services company, dedicated to helping organisations and consumers make commercial and financial decisions with greater confidence and control”. Primarily a credit rating agency similar to DNB.

firms that are under the radar of conventional company lists such as the two mentioned above so finding supporting financials for example for some of these type of firms would be difficult. What the methodology developed does not do is ‘find’ clusters by some fully automated process whereby candidate clusters ‘emerge’ from the search process on the basis that majority word counts equals a certain type of cluster in the manner of a form of keyword based location quotient. The main reasons for this are that the most popular words in the RBKS database are not necessarily those associated directly with certain types of firms and indeed have little to do with industrial activity at all, one of the most popular examples of keyword count being ‘Adobe’.

With regard to the work undertaken on collaborative networks the early attempts to find embedded hyperlinks as a proxy for ‘connections’ between a bounded set of firms were less successful and it was only by extending first the out-link references to other third parties as observed on a site plus externally referenced in-links that changed the picture, particularly for the Subsea sector which was subsequently studied in some detail. The final network diagrams drawn as a result of these exercises have started to give a hint of the power of automated plus manual tracking of internet derived links throughout a population of firms and other organisations.

In summary the larger subsea cohort which was a combination of a proportion of firms derived from conventional snowball search methods and use of the RBKS yielded a far richer corpus of firms and associated information than either method used in isolation. Further the exercise for finding collaborations amongst companies engaged in subsea related activities and their first level collaborators yielded the type of membership and network that fits well with most of the definitions of a cluster as outlined in Appendix 3. Clusters: A Variety of Definitions and in this respect the work contributes to answering the research question.

Contribution to knowledge

The parts that are new are those outlined under What has worked on page 283 onwards but it is the methodology not the final outcome that is deemed important in terms of gaining knowledge. The basic principle of using the internet as a large scale information resource targeted specifically at the finding of industrial clusters and networks does seem to work at the regional level even when using a minority sample of URLs. There seems, at the time of writing, little in the way of prior research that has been targeted directly and on a large scale at the process of augmenting our understanding of what companies actually do; by using the internet as the vehicle for this and what has been undertaken by other investigators has usually been as an adjunct to other forms of research such as machine learning. As a consequence of this the idea of using the internet on a large scale to search for evidence of industrial clustering, whether by finding multiple instances of some commonality of economic activity or by the presence of some form of connections between companies or organisations does not seem to have been used before. This may be for two reasons. Firstly the difficulty of dealing with the large volume of extraneous noise on the internet which tends to obscure the desired information on groups of companies or related organisations operating in some form of agglomeration or milieu. It can be very discouraging for any researcher using the internet and various search tools on a large scale to try and deal with this issue and many of the new search tools available for dealing with it (albeit on a topic not directly related to industrial research) have only become available in recent years. As a result many researchers appear to have pursued a path whereby the information available on the firm has been gleaned through the medium of SICs with the results then having been subject to often highly sophisticated methods of statistical inference to determine the presence or otherwise of industrial clustering. Secondly the temptation to assume that the use of a search engine such as Google Local could find all the companies in a locale that would be required to build a company information database. As discussed in Section 8.4.2 on page 139 onwards this is not an ideal tool because of the difficulty of controlling the search parameters and again researchers following this route may have abandoned such a line of enquiry as not being a practicable proposition.

A further aspect of new knowledge is that concerned with the finding of ‘hidden’ activity. Further as noted by a number of researchers in Sections 3.2, 3.4 and 7.8 regarding problems with SICs, they felt that significant clusters remained hidden or undiscovered because of the shortcomings of SIC systems. As a result of this research we have the ability to circumvent completely the shortcomings of SIC based systems and probe at a deep level the many activities and markets in which individual firms become engaged.

The knowledge base is also strengthened, at a methodological level by an understanding of what does not work and in this respect attempts to fully automate the entire process of data acquisition, translation to information and then knowledge regarding industry in a region are fraught with difficulties, long processing times and inaccuracies. Some of these difficulties may be ameliorated to some degree in the future as new and more efficient programs for data acquisition and subsequent processing become available. In addition it would be expected that web site designers would become more educated to the advantages to be accrued from placing interesting and relevant links on their client's site and reduce reliance on flash based programming aimed at making a visual impact but which adds little else. These advantages would be aimed primarily at raising the profile of any particular website with search engine ranking systems but as it happens additional links both in and out are beneficial to the type of research carried out here.

Contribution to practice

Much has been discussed regarding the limitations of results that are a consequence of not having a nearly complete or a very large sample of regional URLs and the biases that this engenders. However even with these known shortcomings some interesting results have emerged such as the observation of a corpus of firms within the subsea cluster who are involved in one or more other sectors and which with some examination of what these firms do (again using information from the web) we can discern common requirements for specific sets of engineering and project management skills.

In terms of a contribution to the work of practitioners the research here outlines how a number of problems might be addressed for those engaged in trying to understand the make up and organisation of industry within a region and examples of such users might be the RDAs, government agencies and various Consultants. For people charged with regional economic development therefore the method gives a very detailed picture not available by other means of the visibility of even very small firms, how they interact for trade and knowledge exchange and how they may contribute to more than one market segment where the notion of 'related variety' readily becomes apparent [Boschma and Iammarino (2006, 2008)]. An example, already given was of firms engaged in the offshore segment who were seen to be also involved in one or more other areas such as marine, subsea, defence or environmental engineering where a common factor, by observation of this type of firm's internal processes, appeared to be an ability to project manage medium/high complexity engineering design and manufactures for hostile environments. This kind of knowledge, coupled with some intelligent thought regarding such design and manufacturing process in this case is much more finely attuned to what is actually happening to parts of a regional economy and yields more practical (and up to date) results than a wholly SIC based analysis aimed at showing the presence of one or two client predetermined or hoped for clusters. Similarly for foreign direct investment (FDI) agencies acting on behalf of potential inward investors it is possible to answer questions for a regional supply network[41] of the nature:

- Who knows what?
- How do I find and establish my supply chain?
- What is their relationship to others?
- Where are they located?
- Who are hubs and authorities (movers and shakers) in my industry or the ones we did not know existed?
- How will a new entrant 'fit' in the network?

Recommendations for Further Research

During the course of the research many avenues of interest opened up at various points, only some of which could be pursued as they did not always contribute to answering the research question. Other lines of research were promising but for one reason or other as explained in the text had to be curtailed in the interests of overall progress and a further category; that of research carried out here that requires

[41] An opportunity to test this out occurred in early 2008 with a group of Norwegian engineering businesses on an inward trade mission to North East England organised by www.norway2uk.com . One of these companies expressed a particular interest in sealing technology. References to trade directories, the RDA's databases and other sources found only a few companies involved in the sealing business in some way. By contrast a search on the RBKS for words beginning with 'seal' which included, sealant/s, sealing, sealed etc. found 150 organisations. Not all of these were appropriate but the top 50, found in matter of minutes were highly relevant and of value to the company.

development was also identified. Some of the latter issues were outlined above under the sub heading What could work better on page 284.

The first recommendation therefore is for continued development of inter-organisational linking particularly from the point of view of trying to assess the 'worth' or strength of these links. In network theory the notion of 'fitness' of nodes was discussed but it is difficult to carry this forward if all links are ranked the same. The number of links to a 'fit' node (or indeed any node) can be counted as shown in the metrics for degree but some links are more important than others in assessing what is a particularly fit node. In snowball sampling once the appropriate firm has been identified it is relatively simple when counselling expert opinion regarding linkages to ask the expert how important any particular link is from or to their own organisation (even though that opinion may be subject to all the biases that come with such 'expert' opinion). The hands-off internet version of doing the same thing is more difficult as we end up making a judgement on the worth of either an out-link or an in-link derived from the web. When the section on in-links was being written however it was noted that there were two distinct methods of assessing these connections. One was the method adopted here and that was by using straightforward link popularity programs that searched for evidence of in-links to a specified website. Another possible method noted on pages 166 and 205 was that associated with the assessment of 'traffic' to and from a nominated website. This a whole new branch of information mining related to the internet but potentially such investigations, for certain types of firms would hold out the prospect of making an assessment of this elusive parameter of link strength. Although it is by no means certain that such a result could be found it is probably worth further investigation.

A further area for research associated with the topic of linkages is that hinted at in Section 11.2 and concerns measurements of clustering activity based on industrial networks. Once a network such as that shown in Figure 34 has been established and assuming one has some confidence in the robustness of the data then it is relatively easy to use social network analysis software to determine the magnitude of key metrics such as clustering co-efficients. The difficulties arise when trying to compare some supposed cluster (as determined by the metric) either with some other cluster or by comparison with the milieu as a whole. As noted this would be analogous to location quotients that are usually based on employment count. If the employment count associated with an industry or group of industries is greater than the surrounding or national norms the assumption is made that there is an increase in ‘activity’ which is a proxy or an indicator for the presence of an industrial cluster. Here we are proposing a similar thing but using linkages instead of head count. The further assumption would have to be made that such networking activity is proportional to economic activity and not just that some networks ‘talk a lot’ or are well connected because they can be so e.g. digital industries. Midmore, Munday and Roberts (2006) have attempted to do an analysis for industry linkages using regional input-ouput tables in that they established economic ‘linkage indicators’ for sectors in Wales but again comparisons of any metrics are not particularly appropriate. Perhaps of more interest is the work by Casanueva and González (2004) who looked at the interaction of social and information relations in networks of SMEs. Although many works of this type are often qualitative in nature here the authors also undertook a comparison of information interchanges within and between sectors using UCINET. Conversely however the work by Simmie and Sennet (1999) as outlined as early as page 46 indicated that for some of the core metropolitan industries studied there was evidence of clustering but without any noticeable interaction between firms in the manner of the networking studied here and that such innovative firms seemed to be gathered together not so much because they needed or used strong intra-industry networks or linkages, but rather because they were making use of the multiple pick-and-mix possibilities provided by the urbanisation effects of large urban conglomerations.

So using networks as a proxy for clustering effects may be not so straightforward, it is clearly sector dependent and a good starting point for a future research project would be to gain data on <u>all</u> possible intra company linkages for a defined area, establish an overall clustering metric and then search for sets of linkages that had a clustering co-efficient substantially above the ‘average’.

The second line of enquiry that could be pursued with respect to linkages is that associated with non-local networks. When the research question was being framed, for reasons as discussed it was decided to limit the search activity to the N.E. region of England. During the research it became increasingly clear that for the subsea sector for example, companies were for the most part globally oriented but this element was discounted when looking for common associations within the region. It would be interesting to look at global associations found using the same basic methods as for the regional links and to do it for a number of different sectors or types of cluster. The proportion of local/global or the so-called ‘glocal’ activity could then be measured [c.f. Castells M, Ch. 2, 5. (2000)]. Twenty five years ago Granovetter (1973) showed the importance of ‘the strength of weak links’, that is the ability of one dense network to be joined to another remote network by a small number of individual links but the important part played by these ‘connectors’ was out of all proportion to the actual number of occurrences. This observation can be seen on the circle diagram such as Figure 31 but this is only for regional firms and similar forms could be expected when including ‘out of region’ networks.

A third line of investigation that could be carried out would be to increase our understanding of professional communities. As has already been noted the 254 URLs that had words starting with ‘architect’ showed little in the way of inter-regional embedded linking. This at the time was a disappointment as the notion of Communities of Practice has gained currency in recent years and it was felt that a professional grouping such as Architects and their associated service sector would show many linkages. However it may be recalled that when the Subsea segment was first spidered for embedded hyperlinks this sector too showed very few such associations and it was only when observed links and in-links were taken into account that the rich web of interaction became evident. It would therefore be interesting to expose some of these Cs of P to the full treatment with regard to discerning all types of links, to see what emerged.

Final remarks.

The research undertaken here has indicated three main themes.

Firstly it is possible to gain a good understanding of ‘what goes on’ in industry in a defined region by use of the web based search methods developed here. These methods do not of themselves identify industrial clusters directly but should be regarded as a useful adjunct to other more established methods of determining the presence of such clusters. In particular the methodology helps to bridge some of the knowledge gaps in the process of identifying clustering activity either through the identification of candidate firms not easily found by more conventional SIC based searching and also by finding evidence of networking amongst groups of firms related by some association.

Secondly the use of the internet and the world wide web for research of the type undertaken here is fraught with difficulties and the possibilities for error at almost every turn. The reasons for this have been much discussed throughout the thesis but an understanding has been gained of the extent to which such difficulties affect the research and how in such cases the problems might be ameliorated. This then is knowledge that may help others to avoid some of the pitfalls, problematical areas and general false trails found at various points in the research process.

The final theme is that related to future work. During the currency of this research the internet and associated search tools have moved on and there are starting to appear for example ‘scraper’ tools of the type that gave so much trouble in the early part of this investigation. It is reasonable to expect continued developments along the same general theme and thus many of the problems with text and data finding and manipulation which were highlighted in earlier work may be more easily solvable in the near future. In some respects the research here has opened up a number of possibilities for interesting future work, particularly for the ‘hands off’ network scanning approach driven in part by an increase in internet usage, more sophisticated website users and developers and thus the availability of more company websites. All this activity supports the kind of research outlined here and enhances the prospects for web based methods being used for a better understanding of industrial networks and ultimately clusters.

References cited in the Text

Agrawal R, Mamilla H, Srikant R, Toivonen H, and Verkamo A. (1996). Fast Discovery of Association Rules. In In U Fayyad, G Piatetsky-Shapiro, P Smythe and R Uthurusamy (Eds) - Advances in Knowledge and Discovery Mining, AAAI Press/The MIT Press, pp.307-326

Eisenhardt K, Schoonoven C-B. (1990). Organizational growth: linking founding team, strategy, environment, and growth among U.S. semiconductor ventures, 1978-1988. Administrative Science Quarterly , 35, 504 - 29.

Ahonen H, Hemonen O, Kiemettine M and Verkamo A. (1997). Applying Datamining techniques in Data Analysis. In Report C-1997-23,Dept. of |Computer Science, Univ. of Helsinki

Allen P.M. (1997a) Cities and regions as self-organizing systems — models of complexity. Amsterdam, Gordon and Breach Science Publishers

Allen, P. M. (1997b). Cities and Regions as Evolutionary, Complex Systems. Geographical Systems 4, 103-130.

Amit R. and Schoemaker P. (1993). Strategic Assets and Organisational Rent. Strategic Management Journal 14, 33-46

Amit R. and Zott C., (2001). Value Creation in e-business. Strategic Management Journal 22, 493-520

Andrianni P. and Siedlok F. (2005). The Emergence of the Subsea Cluster in the North East. Conference on emergence of Subsea Technology Cluster (STC) in the North East of England. St James' Park, Newcastle upon Tyne NE1 4ST. http://www.renewables-club.com/NOF

Arthur, B. (1989). Competing technologies, increasing returns, and lock-in by historical events. Economic Journal 99: 116-31.

Arthur, B. (1990). ‘Silicon Valley’ locational clusters: When do increasing returns imply Monopoly? Mathematical Social Sciences 19: 235-51.

Arthur, B. (1990). Positive feedbacks in the economy. Scientific American, February, 92-9.

Asheim, B. T. (1997). Industrial districts as ‘learning regions’: A condition for prosperity. European Planning Studies 4 (4): 379-400.

Asheim, B. T., and A. Isaksen. (1997). Location, agglomeration, and innovation: Towards regional innovation systems in Norway? European Planning Studies 5 (3): 299-330.

Assimakopoulos D. (2007). Technological Communities and Networks. Routledge, Advances in Management and Business Studies. ISBN10: 0-415-33480-2

Autio E., Garnsey E. (1997). Early Growth and External Relations in New Technology-Based Firms. Otakaari

Barabasi, A B. (2003). Linked - How everything is connected to Everything Else and what it means for Business, Science and Everyday Life. Penguin Books

Barish O, Knoblock C A, Chen Y-S, Minton S, Philpot A and Shahabi C. (1999). Theaterloc: A Case Study in Information Integration. In IJCAI Workshop on Intelligent Information Integration. Stockholm 1999.

Barney J.B. (1991). Firm resources and sustained competitive advantage. Journal of Management 17, 99-120

Bathelt H, Malmberg A, Maskell, P. (2004). Clusters and Knowledge: Local Buzz, Global Pipelines and the Process of Knowledge Creation. In: Progress in Human Geography. (Vol. 28) pp. 31-56

Becattini, G. (1978). The development of light industry in Tuscany: An interpretation. Economic Notes 3: 107-23.

Becattini, G. (1979). Dal settore industriale al distretto industriale. Alcune considerazione sull’ unit di indagine dell economia industriale. Rivista di Economia e Politica Industrial, no.1. and Becattini, G. (1979). Del “sector” industrial al “districte” industrial, Revista Econòmica de Catalunya, núm 1., 1986 (subsequently translated into English).

Becattini, G., (1989). Sectors and/or districts: some remarks on the conceptual foundations of industrial economics., in: E. Goodman and J. Bamford (Editors), *Small Firms and Industrial Districts in Italy*; Routledge, London and New York, pp. 123-135.

Becattini, G., (1990). The Marshallian industrial district as a socio-economic notion, in Becattini, G., Pyke, F., Sengenberger, W. (eds), *Industrial districts and Inter-firm co-operation in Italy,* International Institute for Labour Studies, Geneva, 37-52

Becattini, G., (1992). The Marshallian industrial district as a socio-economic notion, in G. Becattini, F. Pyke and W. Sengenberger (eds.), Industrial districts and Inter-firm co-operation in Italy, 37-52, International Institute for Labour Studies: Geneva.

Benneworth, P. and Charles, D. (2001). Bridging Cluster Theory and Practice:Learning from the Cluster Policy Cycle. In OECD (Ed.), Innovative Clusters :Drivers of National Innovation Systems (pp. 375-416). Paris: OECD.

Bergman, E. M., and P. Lehner. (1998. 'Regional Industrial Clusters in Austria: Mapping and Documentation for UNIDO Clients,' *IIR working paper, Vienna University of* Economics and Business. Individual clusters also available on UNIDO webpage.

Bergman, E.M. (1998). Industrial Trade Clusters in Action: Seeing Regional Economies Whole, in Steiner, M. (Ed.) Clusters and Regional Specialisation: On Geography, Technology and Networks, London: Pion, pp. 92-110.

Brenner, T. & Weigelt, N. (2001). The Evolution of Industrial Clusters - Simulating Spatial

Brenner, T. (2000). The Evolution of Localised Industrial Clusters: Identifying the Processes of Self-or ganisation. Papers on Economics & Evolution #0011, Max-Planek-Institute Jena.

Briscoe B, Odlyzko A, Tilly B. (2006). Metcalfe's Law is Wrong. Spectrum On-Line. IEEE Spectrum, http://www.spectrum.ieee.org/jul06/4109

Cairncross, F. (1997). The Death of Distance, London: Orion Business Books.

Carlino, G. A. (1979. Increasing returns to scale in metropolitan manufacturing. Journal of Regional Science 19: 363-73.

Casanueva C, González J L G. (2004). Social and Information Relations in Networks of Small and Medium-Sized Firms. M@n@gement, 7(3): 215-238.

Castells M, (2000). The Rise of the Network Society. Blackwell Publishing.

Chakrabarti S., Dom B.E., Gibson D.,Kleinberg J., Kumar R., Raghavan P., Rajagopalan S., Tomkins A. (1999). Mining the Link Structure of the World Wide Web. February 1999. {http://www.cs.cornell.edu/home/kleinber/ieee99.ps.}

Chowdhury, Gobinda G. (1999). Template Mining for Information Extraction from Digital Documents. Library Trends, v48 n1 p182-208 Sum 1999

Crouch, C, Farrell, H. (2001). Great Britain: falling through the holes in the network concept. In C. Crouch, P. Le Gales, C. Trogilia, and H. Voelzkow (eds) Local Production System in Europe: Rise or Demise? Oxford: Oxford University Press: 161-211.

Czamanski S., and L. A. Ablas. (1979). Identification of industrial clusters and complexes: a comparison of methods and findings, Urban Studies 16, 61-80.

Czamanski, S. (1971). Some empirical evidence of the strengths of linkages between groups of industries in urban regional complexes. Papers of the Regional Science Association 27: 137-50.

Dalum, B. (1995). Local and Global Linkages: The Radiocommunications Cluster in Northern Denmark. mimeo, Aalborg University.

Davis G, Yoo M, Baker W.E., (2001). The Small World of the American Corporate Elite. Strategic Organization Vol.1(3) pp301-326

de la Mothe J. and Paquet G. (1998). Local and Regional Systems of Innovation as Learning socio-economies, in de la Mothe J. and Paquet G. (1998) Local and Regional Systems of Innovation. Kluwer Academic Publisher, Norvel, MA, USA, 1-26

Doeringer, P. B., and D. G. Terkla. (1995). Business strategy and cross-industry clusters. Economic Development Quarterly 9 (3): 225-37.

Doeringer, P. B., and D. G. Terkla. (1996). Why do industries cluster? In Business Networks: Prospects for Regional Development, edited by U. H. Staber et al., Berlin, Walter de Gruyter.

Duguid P, Brown J S . (2000). The Social Life of Information. Harvard Business School Press, Bost, MA.

Enright, M. (1996). Regional Clusters and Economic Development: A Research Agenda, in Staber, U., Schaefer, N. and Sharma, B., (Eds.) Business Networks: Prospects for Regional Development, Berlin: Walter de Gruyter,pp. 190- 213.

Ernst D. (1994). Inter-Firms Networks and Market Structure: Driving Forces, Barriers and Patterns of Control, Berkley, CA University of California, BRIE monograph. 1994

Euler L, (1913). Opera Omnia (Collected Works). Berkhauser Verlag AG, Basel, Switzerland

Faloutsos M, Faloutsos P, Faloutsos C. (1999). On Power Law Relationships of the Internet Topology. (ACM SIGGCOM 99,comp) Computer Communications Review 29 (1999);251

Fass C, Ginelli M, Turtle. (1996). Six Degrees of Kevin Bacon. New York: Plume, 1996

Feldman M P, Francis J, Bercovitz J. (2005). Creating a Cluster While Building a Firm: Entrepreneurs and the Formation of Industrial Clusters. Regional Studies, Vol.39.1 February 2005

Feser, E. J. (1998). Enterprises, external economies, and economic development. Journal of Planning Literature 12 (3): 283-302.

Feser E.J. and Bergmann E.M., May (1998). National Industry Cluster Templates: A Framework for Applied Regional Cluster Analysis Regional Studies, 34.1. pp. 1-19

Feser, E.J. and Bergman, E.M. (2000). National Industry Cluster Templates: A Framework for Regional Cluster Analysis, Regional Studies, 34, 1, pp. 1-20.

Formica P.; Mitra J. (2000). Approaches to clustering and market creation in the dot-com economy. Industry and Higher Education, Volume 14, Number 6, 1

December 2000 , pp. 413-423(11)

Freeman, L. C. (1999). Using molecular modeling software in social network analysis. School of Social Sciences, University of California, Irvine. Copy available at eclectic.ss.uci.edu/~lin/chem.html.

Freeman C and Louca F. (2001). As times goes by. From the Industrial Revolution to the Information Revolution. Oxford University Press 2001,

Freitag D. (1998). Information Extraction from HTML: Application of a General Machine Learning Approach. © 1998 American Association for Artificial Intelligence.

Garnsey E. and Heffernan P. (2005). High Technology Clustering through Spin-out and Attraction: The Cambridge Case. Regional Studies Vol. 39, No. 8, November 2005 pp. 1127

Ghani R,. Jones R., Mladenic D., Nigam K., Slaterry S. (1999). Data Mining on Symbolic Knowledge Extracted from the Web. School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

Glasmeier, A. (2000). Economic Geography in Practice: Local Economic Development Policy, pages 559-579 in Clark, G., Feldman, M. and Gertler, M. (Eds.) The Oxford Handbook of Economic Geography, Oxford: Oxford University Press.

Granovetter M.S. (1973). The Strength of Weak Ties. American Journal of Sociology 78, 1973 pp. 1360-1380

Gulati R., Nohria N. and Zaheer A. (2000). Strategic Net-works. Strategic Management Journal - Special issue 21, 203-215

Hearst M. (1999). Untangling Text Mining. In Proceedings of ACL'99: the 37th Annual meeting of the Association for Computational Linguistics1999

Heidenreich, M. (1996). Beyond flexible specialization: The rearrangement of regional production orders in Emilia-Romagna and Baden-Württemberg. European Planning Studies 4 (4): 401-19.

Henry N, Pollard J, and Benneworth P. (2006). Putting Clusters in their Place. In 'Clusters and Regional Development', Ashiem, Cooke and Martin (Eds).

Routledge 2006

Henry N. (2003). Punching above its weight - The Motorsport industry in Northamptonshire. www.campus.ncl.ac.uk/unbs/curds/summary.pdf

Hewings, G. J., G. R. Schindler, P. R. Israilevich, and M. Sonis. (1998). Agglomeration, clustering, and structural change: Interpreting changes in the Chicago regional economy. In Clusters and Regional Specialisation, edited by M. Steiner. London: Pion.

Hite, J M. & Hesterly, W S. (2001). The evolution of firm networks. Strategic Management Journal, 22(3), 275-286

Hoover, E. M. (1937). Location Theory and the Shoe and Leather Industries. Cambridge, MA: Harvard University Press.

Isaksen, A. (1997). Regional clusters and competitiveness: The Norwegian case. European Planning Studies 5 (1): 65-76.

Isard, W. (1956). Location and Space Economy. New York: John Wiley.

Karonski M and Rucinski A. (1997). The Origins of the Random Theory of Graphs, in *The Mathematics of Paul Erdos*, ed. R.L.Graham and J Nesetril. Berlin:Springer, 1997.

Kay J. (1993). Foundations of Corporate Success. Oxford University Press. 1993.

Kelly K. (1995). Out of Control: The Rise of Neo-biological Civilization. ACM SIGCAS Computers and Society Volume 25 , Issue 3. September 1995. ISSN:0095-2737

Kleinberg J. M., (1998). Authoratative Sources in a Hyperlinked Environment. Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998 and IBM Research Report RJ 10076, May 1997, also Journal of ACM

Knoblock A, Minton S, Ambite J, Ashish N, Modi P J, Muslea I, Philpot A, Tejada S. (1998). Modeling Web Sources for Information Integration. AAAI/IAAI 1998: 211-218

Knox P. and Agnew J. (1998). The Geography of the World Economy. Arnold, New

York

Kogut B, and Walker G. (2001). The Small World of Germany and the Durability of National Networks. American Sociological Review 66:317-335

Krackhard, T. D., J. Blythe, and C. Mcgrath. (1994). KrackPlot 3.0: An improved network drawing program, Connections 17, 53-5.

Krackhardt, D., G. Lundberg, and L. O'Rourke. (1993). KrackPlot: A picture's worth a thousand words, Connections 16, 37-47.

Krebs V. (2002). Uncloaking Terrorist Webs. First Monday. http://www.firstmonday.org/issues/issue7_4/krebs/ March 2002

Krugman, P. (1991b). Increasing Returns and Economic Geography. Journal of Political Economy 99, 483-499.

Krugman, P. (1996). Competitiveness: A Dangerous Obsession, Chapter 1, and Myths and Realities of US Competitiveness, Chapter 6, in Pop Internationalism, Cambridge, Mass: MIT Press, pp. 3-24 and pp. 87- 104.

Krugman, P. (1991). Geography and Trade. Cambridge: MIT Press.

Kuah, A. T. H. (2002). Cluster Theory and Practice: Advantage for the Small Business Locating in a Vibrant cluster. Journal of Research in Marketing and Entrepreneurship, Volume Four, Issue 3.

N. Kushmerick, D. Weld and R. Doorenbos. (1997). Wrapper induction for information extraction, IJCAI-97, 1997. http://citeseer.ist.psu.edu/kushmerick97wrapper.html

Lawson, M.; Kemp, N.; Lynch, M. F.; & Chowdhury, G. G. (1996). Automatic extraction of citations from the text of English language patents: An example of template mining. Journal of Information Science, 22(6), 423-436

Lee C.M., Miller W.F., Hancock M.G., Rowen H.S. (Eds.). (2000). The Silicon Valley Edge. Stanford University Press. 2000

Lever, W. F. (1972). Industrial movement, spatial association and functional linkages. Regional Studies 6: 371-84.

Lichtenberg, R. M. (1960). One-tenth of a Nation. Cambridge, MA: Harvard University Press.

Lundvall, B. (1996). National systems of innovation and input-output analysis. In Economic Interdependence and Innovative Activity, by C. DeBresson et al., 356-63. Cheltenham, UK: Edward Elgar.

M. Craven, S. Slattery, and K. Nigam. (1998). First-order learning for web mining. In Proceedings of the 10th European Conference on Machine Learning, Chemnitz, GERMANY, April 1998. Springer Verlag. http://citeseer.ist.psu.edu/craven98firstorder.html

Magee, Gary B. (2005). Review: Industrial Clusters and Regional Business Networks in England, 1750–1970. English Historical Review, Volume 120, Number 487, June 2005, pp. 855-856(2). Oxford University Press.

Maillat, D. (1991). The innovation process and the role of the milieu. In Regions Reconsidered: Economic Networks, Innovation, and Local Development, edited by E. M. Bergman, G. Maier, and F. Todtling, 103-17. London: Mansell.

Malecki, E. J. (1997). Technology and Economic Development. Essex: Addison Wesley Longman.

Malmberg A., Maskell P. (2002). The Elusive Concept of Localisation Economies: Towards Knowledge Based Theory of Spatial Clustering. Environ. Plann. A34, 429-49

Markusen, A. (1998). Sticky Places in Slippery Space, Economic Geography, Economic Geography, 72, p. 293-313.

Marshall (1890). Principles of Economics, London: Macmillan.

Martin, R.L. (1999). The New ‘Geographical Turn’ in Economics: Some Critical Reflections, Cambridge Journal of Economics, 23, pp. 65-91.

Martin, R.L. and Sunley, P (2001). Deconstructing Clusters: Chaotic Concept or Policy Panacea? Regional Studies Association Conference on Regionalising the Knowledge Economy. London.

Meyer-Stamer, J. (1998). Path dependence and regional development: Persistence and change in three industrial clusters in Santa Catarina, Brazil. World Development 26, 1495-1511.

Miller, P. Botham, R. Martin, R. and Moore, B. (2001). Business Clusters in the UK: A First Assessment, London: Department of Trade and Industry.

Miller D., Garnsey E., (2000). Entrepreneurs and Technology Diffusion: How Diffusion Research Can Benefit from a Greater Understanding of Entrepreneurship. Technology in Society 22:445-456

Mitchell T M. (1997). Machine Learning. The McGraw-Hill Companies

Murtha T.P., Lenway S.A. and Hart J.A., (2001). Managing New Industry Creation:Global Knowledge Formation and Entrepreneurship in High Technology. Stanford University Press

O'Donoghue D., Gleave W. (2004). A Note on Methods for Measuring Industrial Agglomeration. Regional Studies, Vol. 38.4, pp.419-427, June 2004

Park, S. O. (1997). Dynamics of new industrial districts and regional economic development, paper presented at the 1997 International Symposium on Industrial Park Development and Management, Taipei, Taiwan.

Park, S. O., and A. Markusen. (1995). Generalizing new industrial districts: A theoretical agenda and an application from a non-Western economy. Environment and Planning A 27 (1): 81-104.

Passiante G. Elia V and Massari T. (2003), Digital Innovation - Innovation Processes in Virtual Clusters and Digital Regions. Imperial College Press.

Passiante G. (2003). Industrial Cluster in the Net-Economy: Emprical Evidence and some Theoretical Approaches in Passiante G. Elia Vand Massari T. (2003). Digital Innovation - Innovation Processes in Virtual Clusters and Digital Regions. Imperial College Press.

Penrose E.T. (1995). The Theory of Growth of the Firm. Basil Blackwell, London

Perry, M. (1999). Clusters Last Stand, Planning Practice and Research, 14, 2, pp. 149-152.

Peteraf M.A. (1993). The cornerstones of competitive advantage: A resource based view. Startegic Management Journal, 14, 179-191

Pierre J.P. (2001). On the Automated Classification of Web Sites. Linkoping Electronic Articles in Computer and Information Science. Vol.6(2001): nr 0

Pollard D. Dec. (2006). "A Whirlwind Tour of Social Networking". Online Information Conference, London. http://www.online-information.co.uk/ol06/index.html

Porter, M. E., (1990). The competitive advantage of nations (MacMillan, London).

Porter, M.E. (1998b) Location, Clusters and the 'New' Microeconomics of Competition, Business Economics, 33, 1, pp. 7-17.

Porter, M.E. (1998c). Clusters and the New Economics of Competitiveness, Harvard Business Review, December, pp. 77-90.

Porter, M.E. (2000). Location, Competition and Economic Development: Local Clusters in the Global Economy, Economic Development Quarterly, 14, 1, pp. 15-31.

Porter, M.E. (2001). Regions and the New Economics of Competition, in Scott, A. (Ed) Global City Regions, Oxford: Blackwell, pp. 139-152.

Porter, M.E. and van Opstal, D. (2001). US Competitiveness 2001: Strengths, Vulnerability and Long-Term Priorities, Washington: Council on Competitiveness (www.compete.org).

Porter,M. E. (2000a). Locations, Clusters and Company Strategy, in Clark, G.L., Feldman, M. and Gertler, M. (Eds) Handbook of Economic Geography, Oxford: Oxford University Press, pp. 253-274.

Prahalad C.K. and Hamel G. (1990). The Core Competence of the Corporation. Harvard Business Review, May-June, 79-91

Quinlan J R. (1990). Learning Logical Definitions from Relations. Machine Learning, 5:239-2666, 1990

Quinlan J R. and Cameron-Jones R M. (1993). FOIL: A Midterm Report. In Proceedings of the European Conference on Machine Learning Pages, pp. 3-20, Vienna 1993

Raven and Pinch, (2003). The British Kit Car Industry: Understanding a `World of Production', European Urban and Regional Studies 2003; 10: 343-354

Richter, C. E. (1969). The impact of industrial linkages on geographic association. Journal of Regional Science 9: 19-28.

Roberts, B., and R. J. Stimson. (1998). Multi-sectoral qualitative analysis: A tool for assessing the competitiveness of regions and formulating strategies for economic development. Annals of Regional Science 32: 469-494.

Roelandt, T. and den Hertog, P. (1999). Cluster Analysis and Cluster- Based Policy Making in OECD Countries: An Introduction to the Theme, Ch 1 in OECD (1999) Boosting Innovation: The Cluster Approach, Paris: OECD, pp. 9-23.

Rosenfeld, S. (1995a). Industrial Strength Strategies: Regional Business Clusters and Public Policy. Washington, DC: Aspen Institute.

Rosenfeld, S. A. (1995b). Overachievers: Business Clusters that Work. Chapel Hill, NC: Regional Technology Strategies, Inc.

Rosenfeld, S. A. (1997). Bringing business clusters into the mainstream of economic development. European Planning Studies 5 (1): 3-23.

Sadler D, (2004). Cluster Evolution, Transformation and the Steel Industry in N.E. England. Regional Studies Vol. 38.1. pp.55-56, February 2004

Saxenian, A. (1994). Regional Advantage: Culture and Competition in Silicon Valley and Route 128. Cambridge, MA: Harvard University Press.

Schumpeter, J. (1934). The Theory of Economic Development. Cambridge, MA: Harvard University Press.

Scott, A. J., and D. Bergman. (1996). The industrial resurgence of California? In Business Networks: Prospects for Regional Development, edited by U. H. Staber et al., Berlin, Walter de Gruyter.

Sexton D., Landstrom H., eds. (1999). The |Blackwell Handbook of Entrepreneuship. Blackwell Handbooks in Management. Oxford; Blackwell.

Shan, W. (1990). An Empirical Analysis of Organizational Strategies by Entrepreneurial High-Technology Firms. Strategic Management Journal, 11, 129-139,

Simmie, J. and Sennett, J. (1999a). Innovation in the London Metropolitan Region, Ch 4 in Hart, D., Simmie, J., Wood, P. and Sennett, J. Inovative Clusters and Competitive Cities in the UK and Europe, Oxford Brookes School of Planning Working Paper 182.

Simmie, J. and Sennett, J. (1999b). Innovative Clusters: Global or Local Linkages? National Institute Economic Review, 170, pp. 87-98.

Slattery S, Craven M. (1998). Combining Statistical and Relational Methods for Learning in Hypertext Domains. In Proceedings of the 8th Intl. Conf. on Inductive Logic Programming (ILP-98)

Smith A. (1776). An Inquiry into the Nature and Causes of the Wealth of Nations

Soderland S, Lehnert W. (1994). Wrap-up: A Trainable Discourse Module for Information Extraction. Journal of Artificial Intelligence Research (JAIR), 2:131-158, 1994

Soderland S, Fisher D, Lehnert W. (1997). Automatically Learned vs Hand Crafted Text Analysis Rules. Technical Report TC-44, Univ. of Massachusets, Amherst, CIIR 1997

Staber U., (2001). The Structure of Networks in Industrial Districts. Int. J. Urban and Reg. Res. 25, 537-53.

Steiner F, (2005). Formation and Early Growth of Business Webs. Physica-Verlag Heidelberg (Springer Science +Business Media).

Steiner, M. (1998). (Ed.) Clusters and Regional Specialisation: On Geography, Technology and Networks, London: Pion.

Stough R., Stimson J, and Roberts R. (1997). Merging quantitative and expert response data in setting regional economic development policy: Methodology and application., presented at the 19th Annual Research Conference of the Association for Public Policy Analysis and Management, November 6-8, Washington D.C.

Swann, G.M.P. and M. Prevezer. (1996). A Comparison of the Dynamics of Industrial Clustering in Computing and Biotechnology', Research Policy, 25,1139-1157. 1996

Teece D.J., Pisano G. and Shuen A. (1997). Dynamic capabilities and strategic. Strategic Management Journal 18, 509-533

Van den Berg, L., Braun, E. and van Winden, W. (2001). Growth Clusters in European Cities: An Integral Approach, Urban Studies, 38, 1, pp.

Vernon R (1960). Metropolis 1985. Harvard University Press, Cambridge, MA

Voyer R. (1997). Knowledge-based Industrial Clustering:International Comparisons. Nordicity Group Ltd. www.idrc.ca/uploads/user-S/10379994410voyerknowledge.doc

Wasserman, S., and K. Faust. (1994). Social Network Analysis. Cambridge: Cambridge University Press.

Watts D J. (1999). Small worlds - The Dynamics of Networks between Order and Randomness. Princeton University Press

Watts D. and Strogatz S. (1998). Collective Dynamics of 'Small-World' Networks, Nature 393(1998): 440-442

Weber, A. (1929). Theory of the Location of Industries. Trans. C. J. Friedrich. Chicago: University of Chicago Press.

Wernerfelt T.B. (1984). A Resource-Based View of the Firm. Strategic Management Journal 5, 171-180

White, H C, (1981). Production Markets as Induced Role Structures, pp. 1-57 in Samuel Leinhardt, ed., Sociological Methodology 1981, San Francisco: Jossey-

Bass,

Williams, J.R. (1998). The Establishment of an Electronically Linked Cluster in the Offshore Industry in North East England. Advances in Information Technologies: The Business Challenge. Conference, Florence, 1998. Roger J., Stanford-Smith B., Kidd P. (Eds). IOS Press Amsterdam

Williamson O E. (1975). Markets and Hierarchies, Analysis and Antitrust Implications: A Study in the Economics of Internal Organisations. Free Press, New York

Williamson, O. (1975). Markets and Hierarchies. New York: The Free Press.

Woolthius R J A K. (1999). Sleeping with the Enemy - Trust, dependence and contracts in interorganisational relationships. Part of a research programme 'Interorganisational relationships and Innovation' at the University of Twente Entrepreneuship Centre. ISBN 90-365-1333-8

Other works consulted but not cited.

Åberg, Y. (1973). Regional productivity differences in Swedish manufacturing. Regional Science and Urban Economics 3: 131-55.

Abbott T, Andrews S. (1990). The Classification of Manufacturing Industries: an Input-Based Clustering of Activity. Paper provided by Center for Economic Studies, U.S. Census Bureau in its series Working Papers with number 90-7.

Acs, Z.J. & Audretsch, D.B. (1991).Innovation and Small Firms. Cambridge: MIT Press.

Almeida, P. and Kogut, B. (1999) Localization of Knowledge and the Mobility of Engineers in Regional Networks, Management Science, 7, pp. 905-917.

Amin, A. (2000) Industrial Districts, Ch 10 in Sheppard, E. and Barnes, T. (eds) A Companion to Economic Geography Oxford: Blackwell, pp. 149-168.

Amin, A. and Cohendet, P. (1999) Learning and Adaptation in Decentralised Business Networks, Environment and Planning D:Society and Space 17, pp. 87-104.

Amin, A. and Thrift, N.J. (1992). Neo-Marshallian Nodes in Global Networks, International Journal of Urban and Regional research, 16, pp. 571-587.

Amin, A., and K. Robins. (1990. The re-emergence of regional economies? The mythical geography of flexible accumulation. Environment and Planning D: Society and Space 8: 7-34.

Amin, A., and K. Robins. (1991). These are not Marshallian times. In Innovation Networks: Spatial Perspectives, edited by R. Camagni, 105-18. London: Belhaven.

Anderson, G. (1994). Industry clustering for economic development. Economic Development Review, Spring, 26-32.

Anselin, L., Varga, A. & Acs, Z. (1997). Local Geographic Spillovers between

University Research and High Technology Innovations. Journal of Urban Economics 42,

Antonelli, C. (1994). Technological Districts, Localized Spillovers and Productivity Growth. The Italian Evidence on Technologcial Externalities in Core Regions, International Review of AppliedEconomics, , pp. 18- 30.

Appold, S. J., and J. P. Gant. (1993). Agglomeration and plant-level technological change. Draft, School of Public Policy and Management, Carnegie Mellon University.

Appold, S.J. (1995). Agglomeration, Inter-organizational Networks, and Competitive Performance in the US Metalworking Sector, Economic Geography, 71, pp. 27-54

Arrow, K. J. (1962). The economic implications of learning by doing. Review of Economic Studies 29: 155-73.

Asheim, B. (2000) .Industrial Districts: The Contributions of Marshall and Beyond, Ch 21 in Clark, G. L., Feldman, M. and Gertler, M. (Eds) TheOxford Handbook of Economic Geography, Oxford: Oxford UniversityPress, pp. 413-431.

Audretsch, D. (1998). Agglomeration and the Location of Innovative Acitivity, Oxford Review of Economic Policy, 14, 2, pp. 18-29.

Audretsch, D. and Feldman, M.; (1996) R&D Spillovers and the Geography of Innovation and ~production, The American Economic Review, 86, 3, pp. 630- 640.

Audretsch, D. B. & Fritsch, M. (1999). The Industry Component of Regional New Firm Formation Processes. Review of Industrial Organization 15, 239-252.

Audretsch, D. B. (1998). Agglomeration and the Location of Innovative Activity. Oxford Review of Economic Policy 14, 18-29.

Audretsch, D., (1995. Innovation, growth and survival, International Journal of Industrial Organisation 13, 44 1-457.

Ballew, P. D., and R. H. Schnorbus. (1994). Realignment in the auto supplier industry: the rippling effects of Big Three restructuring, Economic Perspectives (Federal Reserve Bank of Chicago), Jan/Feb, 2-9.

Baptista, R. (1998). Clusters, Innovation and Growth: A Survey of the Literature, in Swann, G.M.P., Prevezer, M. and Stout, D. (Eds) The Dynamics of Industrial Clustering: International Comparisons in Computing and Biotechnology, Oxford: Oxford University Press, pp.13-51.

Baptista, R. (2000). Do Innovations Diffuse Faster within Geographical Clusters? International Journal of Industrial Organization, 18, pp. 515-535.

Baptista, R. And G.M.P. Swann. (1998). 'Do Firms in Clusters Innovate More ?' Research Policy, 275), 527-542 1998

Baptista, R. And G.M.P. Swann (1999). 'The Dynamics of Firm Growth and Entry in Industrial Clusters: A Comparison of the US and UK Computer Industries, Journal of Evolutionary Economics, 93, 373-399 1999

Baptista, R. and Swann, P. (1999). A Comparison of Clustering Dynamics in US and UK Computer Industries, Journal of Evolutionary Economics, 9, pp. 373-399.

Baptista, R., (1998), Clusters, innovation and growth: a survey of the literature, in: G. M. P. Swann, M. Prevezer and D. Stout, eds., The dynamics of industrial clusters: international comparisons in computing and biotechnology (Oxford University Press, Oxford) 13-51.

Beaudry C. (2000). Entry, Growth and Patenting in Industrial Clusters: A Study of the Aerospace Industry in the UK, Manchester Business School WP 413.

Beaudry C. (1999). Clusters, Growth and the Age of Firms: A Study of Seven European Countries. European Association for Research in Industrial Economics (EARlE), Turin, September 1999

Beaudry C., C. Cook, N. Pandit and G.M.P. Swann. (1999). Clusters, Growth and the Age of Firms: A Study of Three Industries. INLOCO Report 3.3 to EU TSER Programme, October 1999

Beaudry C., S. Breschi, G. Cook, N. Pandit, L. Sanz-Men6ndez and G.M.P. Swann (1999). Report on the Econometric Evaluation of Industrial Clusters and Innovation. INLOCO Report 3.3 to EU TSER Programme, October 1999

Beaudry, C. and G. M. P. Swann, (1998), Clusters, growth and the age of firms: A study of seven European countries, working paper submitted to the European Economic Review.

Beaudry, C., G.A.S. Cook, N.R. Pandit, N. R. and G.M.P. Swann (1998). Industrial Districts and Localised Technological Knowledge: The Dynamics of Clustered SME Networking. Research Report 3.3 (Clusters, growth and the age of firms; A study of three industries: Aerospace, Broadcasting and Financial Services) for the European Community DG XII

Beaudry, C., S. Breschi and G.M.P. Swann, (1998). Statistical and econometric analysis of clustering and innovation, Research Report for the European Community DGXII.

Beeson, P. (1987). Total factor productivity growth and agglomeration economies in manufacturing: (1959-73. Journal of Regional Science 27: 183-99.

Beeson, P. (1990). Sources of the decline of manufacturing in large metropolitan areas. Journal of Urban Economics 28: 71-86.

Beeson, P. E., and S. Husted. (1989). Patterns and determinants of productive efficiency in state manufacturing. Journal of Regional Science 29: 15-28.

Begovi , B. (1991). The Economic Approach to Optimal City Size. Oxford: Pergamon.

Bellandi, M. (1989). The industrial district in Marshall. In Small Firms and Industrial Districts in Italy, edited by E. Goodman and J. Bamford, 136-52. London: Routledge.

Bellandi, M. (1996). On entrepreneurship, region, and the constitution of scale and scope economies. European Planning Studies 4 (4): 421-38.

Belleflamme, P. , Picard, P. and Thisse, J.F. (2000). An Economic Theory of

Regional Clusters, Journal of Urban Economics, 48, 1, pp. 158-184.

Bellini, N. (1996). Italian industrial districts: Evolution and change. European Planning Studies 4 (1): 3-4.

Benjamin R. (1998). Cybercommunities: better than being there?, in Blueprint to the digital economy (ed. by Tapscott D., Lowy A. and Ticoll D.), New York, McGraw Hill

Bergman E. M., E. J. Feser, and S. Sweeney. (1996). Targeting North Carolina Manufacturing: Understanding the State's Economy Through Industrial Cluster Analysis. UNC Institute for Economic Development, Chapel Hill.

Bergman, E. M., and E. J. Feser. (1997). Industrial, Regional or Spatial Clustering? OECD Workshop Position Paper on Cluster Analyses and Cluster-based Policies (Amsterdam, 10-11.10.97). Paris: OECD Industrial Cluster Focus Group.

Bergman, E. M., and E. J. Feser. (2000). Industrial and Regional Clusters: Concepts and Comparative Applications. West Virginia University, The Web Book of Regional Science at www.rri.wvu.edu/WebBook/Bergman-Feser/contents.htm

Bergman, E. M., E. J. Feser, and J. Schare. (1995). Modern Production Practices and Needs: North Carolina's Transportation Equipment Manufacturers, Chapel Hill, NC, UNC Institute for Economic Development.

Bergman, Edward M. (1998). "Regional Economic Coherence and Industrial Trade Clusters," paper presented at International Workshop on Theories of Regional Development, June 14-16, Uddevalla, Sweden.

Bergsman, J., P. Greenston, and R. Healy. (1972). The agglomeration process in urban growth. Urban Studies 9: 263-88.

Bergsman, J., P. Greenston, and R. Healy. (1975). A classification of economic activities based on location patterns. Journal of Urban Economics 2: 1-28.

Best, M. (2001) The New Competitive Advantage: The Renewal of American

Industry, Oxford: OUP.

Best, M. (1990). The New Competition: Institutions of Industrial Restructuring. Cambridge: Polity Press.

Best, M. and Forrant, R. (1996) Creating Industrial Capacity: Pentagon-led Versus Production-led Industrial Policies, pages in J. Michie and J.Grieve Smith (Eds.) Creating Industrial Capacity: Towards Full Employment, Oxford: Oxford University Press.

Bianconi G., Barabasi A-L. (2001). Competition and Multiscaling in Evolving Networks. Europhysics Letters May 2001: 436-442

Birkinshaw, J. and Hood, N. (2000). Characteristics of Foreign Subsidiaries in Industry Clusters, Journal of Business Studies, 31, 1, pp. 141-154.

Boekholt P. (1997). The public sector at arms length or in charge? Towards a typology of cluster policies. Paper presented at OECD Workshop on Cluster Analysis and Cluster Policies, Amsterdam, Netherlands, 9-10 October.

Bopp, R., and P. Gordon. (1977). Agglomeration economies and industrial economic linkages: comment. Journal of Regional Science 17: 125-7.

Borgatti, S.P., Everett, M.G. and Freeman, L.C. (2002). Ucinet for Windows: Software for Social Network Analysis. Harvard, MA: Analytic Technologies. Last revised 10 May 2005. V.6.181

Boyer M.C. (1999) "Cybercities ", New York, Princeton Architectural Press

Braunerhjelm, P., Carlsonn, B. Cetindamar, D. and Johansson, D. (2000) The Old and the New: The Evolution of Polymer and Biomedical Clusters in Ohio and Sweden, Journal of Evolutionary Economics, 10, 5, pp. 417-488.

Breschi S. (2001). CRENOS Conference, Technological Externalities and Spatial Interaction, Cagliari

Breschi, S., (1999), Spatial patterns of innovation, in: A. Gambardella and F. Malerba, (eds.), The organisation of innovative activity in Europe (Cambridge University Press, Cambridge)

Breschi, S. and Lissoni, F. (2001). Localised Knowledge Spillovers versus Innovative Milieux: Knowledge 'Tacitness' Reconsidered, Papers in Regional Science, 80, pp. 255-273.

Brezis, E.S. and P. Krugman, (1993). Technology and the life-cycle of cities, NBER Working Paper, 4561.

Broder A, Kumar R, Maghoul F, Raghavan P, Rajagopalan S, Stata R, Tomkins A, Wiener J. (2000). Graph Structure in the Web. Ninth Intl. World Wide Web Conference Amsterdam, 2000. www9.org/w9cdrom/160/160.html

Brown, R. (2000). Clusters, Supply Chains and Local Embeddeness in Fyrstad, European Urban and Regional Studies, 7, 4, pp. 291-306.

Brusco S. (1996). Trust, social capital and local development: some lessons from the experience of the Italian districts, in Networks of Enterprises and Local Development, Paris, OECD

Brusco, S. (1982). The Emilian model: Productive decentralisation and social integration. Cambridge Journal of Economics 6: 167-84.

Brusco, S. (1986). Small firms and industrial districts: The experience of Italy. In New Firms and Regional Development in Europe, edited by D. Keeble and E. Wever, 184-202. Beckenham, Kent: Croom Helm.

Buckley, P.J., Pass, C. and Prescott, K. (1988). Measures of International Competitiveness: A Critical Survey, Journal of Marketing Management, 4, pp. 175-200.

Burt, D. N. (1989). Managing suppliers up to speed. Harvard Business Review 67: 127-35.

Caniels, M. C. J. And Verspagen, B. (1999). Spatial Distance in a Technology Gap Model. Working Paper 99.10, Eindhoven Centre for Innovation Studies.

Cappellin R. (1997). "From an European regional policy to an European territorial policy: the role of cities and urban policies", the Regional Science Association 37th European Congress proceedings, Rome, august

Carlino, G. A. (1980). Contrasts in agglomeration: New York and Pittsburgh reconsidered. Urban Studies 17: 343-51.

Carlino, G. A. (1982). Manufacturing agglomeration economies as returns to scale, a production function approach. Papers and Proceedings of the Regional Science Association 50: 95-108.

Carlino, G. A. (1987). Comparisons of agglomeration: or what Chinitz really said: a reply. Urban Studies 24: 75-6.

Carlsson, B. (1996). Small business, flexible technology, and industrial dynamics. In Small Business in the Modern Economy, edited by P. H. Admiraal, 63-125. Oxford: Blackwell.

Carrillo, J. (1995). Flexible production in the auto sector: Industrial reorganization at Ford-Mexico. World Development 23: 87-101.

Castilla E J, Hoyku Hwang, Granovetter E, Granovetter M. (2000). Social Networks in Silicon Valley. In: Lee C.M., Miller W.F., Hancock M.G., Rowen H.S. (Eds.). The Silicon Valley Edge. Stanford University Press. 2000

Cawthorne, P. M. (1995). Of networks and markets: The rise and rise of a South Indian town, the example of Tiruppur's cotton knitwear industry. World Development 23: 43-56.

Chinitz, B. (1961). Contrasts in agglomeration: New York and Pittsburgh. American Economic Review 51 (2): 279-89.

Christopherson, S., and M. Storper. (1989). The effects of flexible specialization on industrial politics and the labor market: The motion picture industry. Industrial and Labor Relations Review 42: 331-47.

Coase, R. H. (1937). The nature of the firm. Economica 4: 386-405.

Coe, N. and Townsend, A.R. (1998). Debunking the Myth of Localised Agglomerations: The Development of the a Regionalized Mode of Service Growth in South East England, Transactions of the Institute of British Geographers, 23, pp. 385-404.

Cohen, S. and Fields, G. (1999). Social Capital and Capital Gains in Silicon Valley, California Management Review, 41, 2, pp. 108-130.

Cook, G. A. S. and Pandit, N. R. (2001). Technology, clustering and small firms: Insights from the British broadcasting industry. Ninth International High Technology Small Firms Conference, Manchester Business School, June 2001

Cooke, P. and Morgan K. (1998). The Associational Economy, Oxford: Oxford University Press.

Cortwright, J., and A. Reamer. (1998). Socioeconomic Data for Understanding your Regional Economy. Washington, DC: Economic Development Administration.

Council on Competitiveness (2001). US Competitiveness 2001: Strengths, Vulnerability and Long-Term Priorities (www.compete.org)

Coyle, D. (1997). The Weightless World: Strategies for Managing the Digital Economy, London: Capstone.

Coyle, D. (2001). Paradoxes of Prosperity: Why the New Capitalism Benefits All, London: Texere Publishing.

Crevoisier O, (1993), "Spatial shifts and the emergence of innovative milieux: the case of the Jura region between 1960 and 1990" Environment and Planning C: Government and Policy 11(4) 419 – 430

Crouch, C., Le Gales, P., Trogilia, C. and Voelzkow, H. (2001). Local Production System in Europe: Rise or Demise? Oxford: Oxford University Press.

Czamanski, S. (1974). Study of Clustering of Industries. Halifax, Nova Scotia: Institute of Public Affairs.

Czamanski, S. (1976). Study of Spatial Industrial Complexes. Halifax, Nova Scotia: Institute of Public Affairs.

Czamanski, S., and L. A. de Ablas. (1979). Identification of industrial clusters and complexes: a comparison of methods and findings. Urban Studies 16: 61-80.

Dahmén, E. (1984). Schumpeterian dynamics. Journal of Economic Behavior and Organizations 5: 25-34.

Dahmén, E. (1988). 'Development blocks' in industrial economics. Scandinavian Economic History Review 36: 3-14.

Darwent, D. (1969). Growth poles and growth centres in regional planning: a review. Environment and Planning 1: 5-31.

David, P. A. and J. L. Rosenbloom, (1990), Marshallian factor market externalities and the dynamics of industrial localisation, Journal of Urban Economics 28, 349-370.

Davidson, K. M. (1992). How should the US encourage innovation? Journal of Business Strategy, March/April, 58-61.

DeBerranger, P. and Meldrum, M.C.R. (2000). The Development of Intelligent Local Clusters to Increase Global Competitiveness and Local Cohesion: The Case of Small Businesses in the Creative Industries, Urban Studies, 37, 10, pp. 1827-1836.

DeBresson, C., and X. Hu. (1997). Techniques to identify innovative clusters: A method and 8 instruments. Paper presented at OECD Workshop on Cluster Analysis and Cluster Policies, Amsterdam, Netherlands, 9-10 October.

Del Ottati, G. (1994). Cooperation and Competition in the Industrial District as an Organisation Model. European Planning Studies 2, 463-483.

Dieperink, H., and P. Nijkamp. (1988). Innovative behaviour, agglomeration economies, and R & D infrastructure. Empirical Economics 13: 37-57.

Dodge M. and Kitchen R. (2001), Mapping Cyberspace. Routledge, London, ISBN 0-415-19884-4

Doeringer P.B. and Terka, D.G. (1996). Why Do Industries Cluster? In Staber, U., Schaefer, N. and Sharma, B., (Eds.) Business Networks: Prospects for Regional Development, Berlin: Walter de Gruyter, pp. 175-189.

Drejer, I., F. S. Kristensen, and K. Laursen. (1997). 'Studies of Clusters as a Basis for Industrial and Technology Policy in the Danish Economy.' OECD Workshop Position Paper on Cluster Analyses and Cluster-based Policies (Amsterdam, 10-11.10.97). Paris: OECD Industrial Cluster Focus Group.

Driffield N, Munday M and Roberts A. (2002). Foreign Direct Investment, Transaction Linkages and the performance of the domestic sector. International Journal of the Economics of Business 9, 335-351

Duffy, N. E. (1988). Returns-to-scale behavior and manufacturing agglomeration economies in U.S. urban areas. Review of Regional Studies 18: 47-54.

Dumais, D., G. Ellison, and E. L. Glaeser. (1997). ‘Geographic Concentration as a Dynamic Process,’ NBER Working Paper No. 6270, Cambridge, MA.

Edel, M. (1972). Land values and the costs of urban congestion: measurement and distribution. In Political Economy of Environment: Problem of Method, 61-90. The Hague: Mouton.

Ellison, G. and Glaeser, E.L. (1997). Geographic Concentration in US Manufacturing Industries: A Dartboard Approach, The Journal of Political Economy, 105, pp. 889-927.

El-Shakhs, S. (1972). Development, primacy and systems of cities. Journal of Developing Areas 7: 11-35.

Enright, M. and Ffowcs-Williams, I. (2001). Local Partnership, Clusters and SME Globalisation, OECD Workshop paper, www.oecd.org.

Evans, A. W. (1986). Comparisons of agglomeration: or what Chinitz really said. Urban Studies 23: 387-9.

Executive Office of the President Office of Management and Budget, Statistics Canada, Statistical Office of the European Communities (n.d.). International Concordance between the Industrial Classifications of The United Nations (ISIC Rev.3), and Canada (1980 SIC), The European Union (NACE Rev.1), and The United States (1987 SIC). Washington, DC: U.S. Bureau of the

Census.

Feldman, M. (2000). Location and Innovation: The New Economic Geography of Innovation, Spillovers, and Agglomeration, in Clark, G.L., Feldman, M. and Gertler, M. (Eds.) Oxford Handbook of Economic Geography, Oxford: Oxford University Press, pp. 373-394.

Feldman, M.P., (1994). The geography of innovation (Kiuwer Academic Publishers, Dordrecht).

Feser, E. J., and S. H. Sweeney. (1999). A test for spatio-economic clustering. Journal of Geographical Systems.

Feser, E.J. (1998). Old and New Theories of Industry Clusters, in Steiner, M. (1998) (Ed.) Clusters and Regional Specialisation: On Geography, Technology and Networks, London: Pion, pp. 18-40.

Fine, C. H., R. St. Clair, J. C. Lafrance, and D. Hillebrand. (1996). The U.S. Automobile Manufacturing Industry. U.S. Department of Commerce, Office of Technology Policy, Washington, DC.

Fisher, E. and Reuben, R. (2000). Industrial Clusters and SME Promotion in Developing Countries, Commonwealth Trade Enterprise Paper Number 3, London: Commonwealth Secretariat.

Florence, P. S. (1948). Investment, Location, and Size of Plant. London: Cambridge University Press.

Fogarty, M. S., and G. A. Garofalo. (1978). An exploration of the real productivity effects of cities. Review of Regional Studies 8: 65-82.

Fogarty, M. S., and G. A. Garofalo. (1988). Urban spatial structure and productivity growth in the manufacturing sector of cities. Journal of Urban Economics 23: 60-70.

Foss, N. J., and C. Knudsen, eds. (1996). Towards a Competence Theory of the Firm. London: Routledge.

Freeman C and Louçã F. (2001). As Time Goes By - From the Industrial Revolutions to the Information Revolution. Oxford University Press, USA, May 2001

Fritz, O. M., H. Mahringer, and M. T. Valderrama. (1998). A risk-oriented analysis of regional clusters. In Clusters and Regional Specialization, edited by M. Steiner, 181-91. London: Pion.

Fujita, M. Krugman, P. and Venables, A. (2000). The Spatial Economy: Cities Regions and International Trade, Cambridge, Mass: MIT Press.

Gertler, M. S. (1988). The limits to flexibility: comments on the post-Fordist vision of production and its geography. Trans. Institute of British Geographers 13: 419-32.

Gertler, M. S. (1993). Implementing advanced manufacturing technologies in mature industrial regions: Towards a social model of technology production. Regional Studies 27: 665-80.

Gibson D., Kleinburg J., Raghavan P. (1998). Inferring Web Communities from Link Topology. Proc. Of the 9th ACM Conference on Hypertext and Hypermedia 1998.

Gilmour, J. M. (1974). External economies of scale, inter-industrial linkages and decision making in manufacturing. In Spatial Perspectives on Industrial Organization and Decisionmaking, edited by F. E. I. Hamilton, 335-62. London: John Wiley and Sons.

Gladwell M. (2000). The Tipping Point. New York:Little Brown 2000

Glaeser, E. L. (1994). Cities, information, and economic growth. Cityscape 1: 9-47.

Glaeser, E. L., J. A. Scheinkman, and A. Shleifer. (1995). Economic growth in a cross-section of cities. Journal of Monetary Economics 36: 117-43.

Glaeser, E., Kallal, H., Scheinkman and Shleifer, A. (1992). Growth in Cities, Journal of Political Economy, 100, 6, pp. 1126- 1152.

Gold, B. (198). Changing perspectives on size, scale, and returns: an interpretive essay. Journal of Economic Literature 19: 5-33.

Goldstein, G. S., and T. J. Gronberg. (1984). Economies of scope and economies of agglomeration. Journal of Urban Economics 16: 91-104.

Gollub, J., ET. AL. (1997). Cluster-based economic development: A key to regional competitiveness–case studies. Springfield, VA, National Technical Information Service PB98-117088.

Gordon, I.R. and McCann, P. (2000). Industrial Clusters: Complexes, Agglomeration and/or Social Networks? Urban Studies, 37, 3, pp. 513-532.

Gray, I. (1998) .False Dawn: The Delusions of Global Capitalism, London:

Gray, M., E. Golob, and A. Markusen. (1996). Big firms, long arms, wide shoulders: The 'Hub-and-Spoke' industrial district in the Seattle region. Regional Studies 30: 651-66.

Greytak, D., and P. Blackley. (1985). Labor productivity and local industry size: further issues. Southern Economic Journal 51: 1121-19.

Griliches, Z. (1992). The search for R&D spillovers. Scandinavian Journal of Economics 94 (Supplement): S29-S47.

Grossman, G. and F. Helpman, (1992). Innovation and growth in the global economy (MIT Press, Cambridge, Mass.)

Guilli, A., Signorini, A. (2006) . The Indexable Web is more than 11.5 billion pages. http://www.cs.uiowa.edu/~asignori/web-size/

Guthrie, J. A. (1955). Economies of scale and regional development. Papers and Proceedings of the Regional Science Association 1: J1-J10.

Hansen, E. R. (1990). Agglomeration economies and industrial decentralization: the wage-productivity trade-offs. Journal of Urban Economics 28: 140-59.

Hansen, N. (1991). Factories in Danish fields: How high-wage, flexible production has succeeded in peripheral Jutland. International Regional Science Review 14: 109-32.

Hanson, G. (2000). Firms, Workers, and the Geographic Concentration of Economic Activity, in Clark, G.L., Feldman, M. and Gertler, M. (Eds.) Oxford Handbook of Economic Geography, Oxford: Oxford University Press, pp. 477-494.

Harfield, T (1998). Strategic Management and Michael Porter: A Postmodern Reading, Electronic Journal of Radical Organisation, 4, 1.

Harreld J. B. (1998). Building smarter, faster Organizations, in Blueprint to the digital economy (ed. by Tapscott D., Lowy A. and Ticoll D.), New York, McGraw Hill

Harrison, B, Kelley, M. and Gant, J. (1996). Innovative Firm Behaviour and Local Milieu: Exploring the Intersection of Agglomeration, Firm Effects, and Technological Change, Economic Geography, 72, pp. 233- 258.

Harrison, B. (1992). Industrial Districts: Old Wine in New Bottles? Regional Studies, 26, pp. 469-483.

Hassink, R. (1997). Localised Industrial Learning and Innovation Policies, European Planning Studies, 5, pp. 279-282.

Held, J. R. (1996). Clusters as an Economic Development Tool: Beyond the Pitfalls, Economic Development Quarterly, 10, 3, pp. 249-261.

Helper S. R. (1994). Three steps forward, two steps back in automotive supplier relations, Technovation 14, 633-40.

Helper, S. R. (1991). Strategy and irreversibility in supplier relations: The case of the U.S. automobile industry. Business History Review 65 (4): 781-824.

Helper, S. R. (1994). Three steps forward, two steps back in automotive supplier relations. Technovation 14: 633-40.

Henderson, J. V. (1986). Efficiency of resource usage and city size. Journal of Urban Economics 19: 47-70.

Henderson, J. V. (1996). Ways to think about urban concentration: Neoclassical urban systems versus the new economic geography. International Regional

Science Review 19: 31-6.

Henderson, V., A. Kuncoro, and M. Turner. (1995). Industrial development in cities. Journal of Political Economy 103: 1067-85.

Higgins, B. (1983). From growth poles to systems of interactions in space. Growth and Change 14: 3-13.

Higgins, B., and D. J. Savoie. (1995). Regional Development Theories and their Application. New Brunswick: Transaction.

Hill, E. W., and J. Brennan. (1998). A methodology for identifying the drivers of industrial clusters: The foundation of regional competitive advantage, forthcoming, Economic Development Quarterly. Vol. 14, No. 1, 65-96 (2000)

Hill, J. and Naroff, J.L. (1984). The Effect of Location on the Performance of High Technology Firms, Financial Management, Spring, pp. 27-36.

Hirschman, A. O. (1958). The Strategy of Economic Development. New Haven: Yale University Press.

Hirschman, A. O. (1987). Linkages. In The New Palgrave Dictionary of Economics, Volume 3, edited by J. Eatwell, M. Milgate, and P. Newman, 206-10. London: Macmillan.

Holmes, J. (1986). The organization and locational structure of production subcontracting. In Production, Work, Territory, edited by A. J. Scott and M. Storper, 80-106. Boston: Allen and Unwin.

Home Page for Leigh Tesfatsion. (to January 2008) http://www.econ.iastate.edu/tesfatsi/

Hotelling, H., (1929). The stability of competition, Economic Journal 39, 41-57.

Howells, J., (1984). The location of research and development: some observations and evidence from Britain, regional Studies 18, 13-29.

Howells, J., (1990). The location and organisation of research and development:

new horizons, Research Policy 19, 133-46.

Hudson, R. (1999). The Learning Economy, the Learning Firm and the Learning Region: A Sympathetic Critique of the Limits to Learning, European Urban and Regional Studies, 6, pp. 59-72.

Huey J. (1996). Twelve themes of the new economy, in The Digital Economy (ed. by Tapscott D.) New York, McGraw Hill

Humphrey, J. (1995). Introduction. World Development 23 (1): 1-7.

Hyun J. (1994). Buyer-supplier relations in the European automobile component industry, Long Range Planning 27, 66-75.

Innovation:The Cluster Approach, Paris: OECD, (1999). pp. 127-153. Also in ISBN: 9789264170803

Isard, W., *et al.* (1998). Methods of Interregional and Regional Analysis. Aldershot: Ashgate.

Isserman, A. M. (1996). It’s obvious, it’s wrong, anyway they said it years ago? International Regional Science Review 19: 37-48.

Jacobs, D., and A-P de Man. (1996). Clusters, industrial policy and firm strategy: A menu approach. Technology Analysis and Strategic Management 8 (4): 425-37.

Jacobs, D., and M. W. de Jong. (1992. Industrial clusters and the competitiveness of The Netherlands. De Economist 140 (2): 233-52.

Jacobs, J. (1969. The Economy of Cities. New York: Random House.

Jacobson, D., and B. Andréosso-O’Callaghan. (1996). Industrial Economics and Organization: A European Perspective. Berkshire: McGraw Hill.

Jaffe, A. B., M. Trajtenberg and R. Henderson, (1993). Geographic localization of knowledge spillovers as evidenced by patent citations, Quarterly Journal of Economics 108, 577-598. Krugman, P., 1991, Geography and Trade (The MIT Press, Cambridge, Mass.)

Juoro, U. (1989). The determinants of agglomeration economies in Indonesia and the Philippines. Philippine Review of Economics and Business 26: 141-71.

Jurvetson, St. (2000). Changing Everything. The Internet Revolution and Silicon Valley”, in: Lee C.M., Miller W.F., Hancock M.G., Rowen H.S. (Eds.). The Silicon Valley Edge. Stanford University Press. 2000

Justman, M. (1995). Infrastructure, growth and the dimensions of industrial policy. Review of Economic Studies 62: 131-57.

Kalakota R., Lowy A. and Ticoll D. (1998). Joined at the bit: the emergence of the e-business community”, in Blueprint to the digital economy (ed. by Tapscott D., Lowy A. and Ticoll D.), New York, McGraw Hill

Kanemoto, Y. (1990). Optimal cities with indivisibility in production and interactions between firms. Journal of Urban Economics 27: 46-59.

Karaska, G. J. (1969). Manufacturing linkages and the Philadelphia economy: some evidence of external agglomeration forces. Geographical Analysis 1: 354-69.

Karonski M and Rucinski A. (1997). The Origins of the Random Theory of Graphs, in The Mathematics of Paul Erdos, ed. R.L.Graham and J Nesetril. Berlin:Springer, 1997.

Kaufman, A., Gittell, R., Merenda, M., Naumes, W., and C. Wood. (1994). Porter’s model for geographic competitive advantage: The case of New Hampshire. Economic Development Quarterly 8 (1): 43-66.

Kawashima, T. (1975. Urban agglomeration economies in manufacturing industries. Papers and Proceedings of Regional Science Association 43: 157-75.

Ke, S. (1995). Beyond Capital and Labor: The Contributions of Technology and Regional Milieu to Production and Productivity Growth. New York: Garland.

Keeble, D. and Wilkinson, F. (2000). (Eds.) High-Technology Clusters, Networking and Collective Learning in Europe, Aldershot: Ashgate.

Keeble, D. E. (1969). Local linkage and manufacturing growth in Outer London. Town Planning Review 40: 163-88.

Keeble, D.E and Nachum, L. (2002). Why do Business Service Firms Cluster? Small Consultancies, Clustering and Decentralisation in London and Southern England, Transactions of the Institute of British Geographers, (forthcoming)

Keller, W. (1997). Trade and the Transmission of Technology. NBER Working Paper No. 6113, Cambridge, MA.

Khoussainov R., O'Meara T., and Patel A. Adaptive distributed search and advertising for WWW, in: N. Callaos et al. (ed.), Proceedings of the 5th World Multi-Conference on Systemics, Cybernetics and Informatics (SCI 2001), vol. 5, pp. 73{78, Orlando, Florida, USA 2001.

Kirkpatrick, D. A., and K. Gavaghan. (1996). North Carolina Environmental Business Study. Prepared by Kirkworks for the North Carolina Environmental Technologies Consortium.

Klier, T. H. (1994). The impact of lean manufacturing on sourcing relationships. Economic Perspectives (Journal of the Federal Reserve Bank of Chicago), July/August, 8-18.

Knudsen, C. (1996). The competence perspective: A historical review. In Towards a Competence Theory of the Firm, edited by N. J. Foss and C. Knudsen, 13-37. London: Routledge.

Kobrin S.J. (1997). The Architecture of Globalization. In Government, Globalization and International Business (ed. by Dunning J. H), New York. Oxford University Press

Koch, R. (1998). The 80/20 Principle - The Secret to Success by Achieving More with Less. New York: Currency, 1998.

Kohn, M., and N. Marion. (1992). The implications of knowledge-based growth for the optimality of open capital markets. Canadian Journal of Economics 25: 865-83.

Krugman, P. (1994). Productivity and Competitiveness, Appendix to Chapter 10, in Peddling Prosperity, New York, W.W. Norton, pp. 268-280.

Krugman, P. (1990). Rethinking International Trade. Cambridge, MA: MIT Press.

Lagendijk, A., and D. Charles. (1997). Clustering as new growth strategy for regional economies? A discussion of new forms of regional industrial policy in the UK. Paper presented at OECD Workshop on Cluster Analysis and Cluster Policies, Amsterdam, Netherlands, 9-10 October.

Lamming, R. (1993). Beyond Partnership: Strategies for Innovation and Lean Supply. Prentice Hall, Hemel Hempstead.

Lawson, C (1997). Territorial Clustering and High-Technology Innovation: From

Lawson, C. and Lorenz, E. (1999). Collective Learning, Tacit Knowledge and Regional Innovative Capacity, Regional Studies, 33, 4, p305.

Lazerson, M. (1988). Organisational growth of small firms: An outcome of markets and hierarchies? American Sociological Review 53: 330-42.

Lazerson, M. and Lorenzoni, G. (1999). The Firms that Feed Industrial Districts: A Return to the Italian Source, Industrial and Corporate Change, Vol. 8, Number 2, pp. 235-266.

Leamer, E. and Storper, M. (2001). The Economic Geography of the Internet Age, Working Paper 8450, Washington: National Bureau of Economic Research.

Leitch S. (2006). Leitch Review of Skills. Prosperity for all in the Global Economy – World Class Skills. HMSO, December 2006

Lessig L. (1999). Code and other Laws of Cyberspace. New York:Basic Books, (1999)

Lever, W. F. (1974). Manufacturing linkages and the search for supplies and markets. In Spatial Perspectives on Industrial Organisation and Decision Making, edited by F. E. I. Hamilton. Chichester, Sussex: John Wiley.

Lipnack J., Stamps J. (1997). Virtual teams, New York, John Wiley & Sons Inc.

Liston, C. (1997). Knitting Mills in North Carolina and Virginia. pp. 150-174, in Regional Technology Strategies (1997).

Loasby, B. (1998). Industrial Districts as Knowledge Communities, in M. Bellet and C. L'Harmet (eds) Industry, Space and Competition: The Contributions of Economists of the Past, Cheltenham: Edward Elgar, pp. 70-85.

Lovering, J. (1999). Theory led by Policy? The Inadequacies of the 'New Regionalism'. International Journal of Urban and Economic Research, 23, pp. 379-395.

Lucas, R. E., Jr. (1988). On the mechanics of economic development. Journal of Monetary Economics 22 (1): 3-42.

Lundequist, P. (2001). Innovative Clusters – Practical Lessons from Regional Cluster Building in Sweden. In Implementing Clusters in Practice – Academic and Practical Perspectives, CURDS in Association with RSA, Newcastle.

Lundvall B. A. (1996). The social dimension of the Learning Economy. Druid Working Paper No. 96-1

Lundvall, B. (1999). From Fordism to the globalising learning economy– implications for innovation policy. Paper presented at the International Seminar on Learning Economy: Innovation-Qualification-Employment, Renner Institut, Vienna, June 21, 1999.

Maillat D. (1996). Regional productive systems and innovative milieus. In Networks of enterprises and local development, Paris, OECD

Maillat, D. (1998). From the Industrial District to the Innovative Milieu: Contribution to an Analysis of Territorialised Productive Organisations. Recherches Economiques de Louvain 64, 111-129.

Maillat, D., and J. Y. Vasserot. (1988). Economic and territorial conditions for indigenous revival in Europe's industrial regions. In High Technology Industry and Innovative Environments, edited by P. Aydalot and D. Keeble, 163-83. Andover: Routledge.

Malerba F. (1993). The National System of Innovation: Italy. In National Innovation System: a comparative analysis, (edited by R.R. Nelson) , New

York, Oxford University Press

Malizia, E. E., and E. J. Feser. (1998). Understanding Local Economic Development. Forthcoming, Rutgers University Press.

Malmberg, A., (1996). Industrial Geography: Agglomeration and Local Milieu, Progress in Human Geography, 20, 3, pp. 392-403.

Malmberg A, Power D. (2006). True Clusters. A Severe Case of Conceptual Headache. In Asheim, B., Cooke, P. & Martin, R. (eds) Clusters and Regional Development. London: Routledge (with D. Power)

Malmberg, A., and P. Maskell. (1997). Towards an explanation of regional specialization and industry agglomeration. European Planning Studies 5 (1): 25-41.

Marceau, J. (1997). The disappearing trick: Clusters in the Australian economy. Paper presented at OECD Workshop on Cluster Analysis and Cluster Policies, Amsterdam, Netherlands, 9-10 October.

Marcus, M. (1965). Agglomeration economies: a suggested approach. Land Economics 41: 279-84.

Markusen, A. (1996). Interaction between regional and industrial policies: Evidence from four countries. International Regional Science Review 19: 49-77.

Markusen, A. (1996). Sticky places in slippery space: A typology of industrial districts. Economic Geography 72: 293-313.

Markusen, A., and S. O. Park. (1993). The state as industrial locator and district builder: the case of Changwon, South Korea, Economic Geography 69, 157-81.

Marshall, A. (1961). Principles of Economics: An Introductory Volume. Ninth (Variorum) Edition (1st Edition 1890). London: Macmillan.

Martin, R. (1990). Flexible futures and post-Fordist places. Environment and Planning A 22: 1276-80.

Martin, R., and P. Sunley. (1996). Paul Krugman's geographical economics and its implications for regional development theory: A critical assessment. Economic Geography 72 (3): 259-92.

Martin, R.L. and Sunley, P. (1998). Slow Convergence? The New Endogenous Growth

May, W., Mason, C. and Pinch, S. (2001). Explaining Industrial Agglomeration: The Case of the British High-Fidelity Industry, Geoforum, 32, 3, pp. 363-376.

Mera, K. (1973). On urban agglomeration and economic efficiency. Economic Development and Cultural Change 21: 309-37.

Meyer, D. R. (1977). Agglomeration economies and urban-industrial growth: Clarification and review of concepts. Regional Science Perspectives 7: 80-91.

Meyer-Stamer, J. (1995). Micro-level innovations and competitiveness, World Development 23, 143-8.

Midmore P, Munday M, Roberts A. (2006). Assessing Industry Linkages Using Regional Input-Output Tables. Regional Studies, May 2006, Vol. 40 No.3

Miller R. E., and P. D. Blair. (1985). Input-Output Analysis: Foundations and Extensions. Prentice Hall, Englewood Cliffs, NJ.

Mitra, J and Matlay, H. (2000). Thematic Clustering: connecting organizational learning in small and medium sized businesses. Industry and Higher Education Journal, Vol. 14, no.6, December, 2000

Moomaw, R. L. (1981). Productive efficiency and region. Southern Economic Journal 48: 344-57.

Moomaw, R. L. (1981. Productivity and city size, a critique of the evidence. Quarterly Journal of Economics 96: 675-88.

Moomaw, R. L. (1983). Spatial productivity variations in manufacturing: a critical survey of cross-sectional analysis. International Regional Science Review 8: 1-22.

Moomaw, R. L. (1985). Firm location and city size, reduced productivity advantages as a factor in the decline of manufacturing in urban areas. Journal of Urban Economics 17: 73-89.

Moomaw, R. L. (1986). Have changes in localization economies been responsible for declining productivity advantages in large cities? Journal of Regional Science 26: 19-32.

Moomaw, R. L. (1988). Agglomeration economies: localization or urbanization? Urban Studies 25: 150-61.

Moomaw, R. L., and M. Williams. (1991). Total factor productivity in manufacturing: further evidence from the states. Journal of Regional Science 31: 17-34.

Morfessis, I. T. (1994). A cluster-analytic approach to identifying and developing state target industries: The case of Arizona. Economic Development Review, Spring, 33-7.

Murray, F. (1987). Flexible specialisation in the Third Italy" Capital and Class 33: 84-95.

Myrdal, G. (1957). Economic Theory and Underdeveloped Regions. New York: Harper and Row.

Nachum, L. and Keeble, D.E. (1999). Neo-Marshallian nodes, Global Networks and Firm Competitiveness: The Media Cluster of Central London, Working Paper 158, ESRC Centre for Business Research, University of Cambridge,.

Nakamura, R. (1985). Agglomeration economies in urban manufacturing industries, a case of Japanese cities. Journal of Urban Economics 17: 108-24.

Nelson, R. R. (1988). Institutions supporting technical change in the United States. In Technical Change and Economic Theory, edited by G. Dosi et al., pp. 312-29. London: Pinter.

Nelson, R. R. (1993). National Innovation Systems: A Comparative Study. New York: Oxford University Press.

Newman, R. G. (1989). Single sourcing: Short-term savings versus long-term problems. Journal of Purchasing and Materials Management 25: 20-25.

Nicholson, N. (1978). Differences in industrial production efficiency between urban and rural markets. Urban Studies 15: 91-5.

Nourse, H. O. (1968). Regional Economics: A Study of the Economic Structure, Stability, and Growth of Regions. New York: McGraw Hill.

Ó hUallacháin, B. (1989). Agglomeration of services in American metropolitan areas. Growth and Change 20: 34-49.

Ó hUallacháin, B., and M. A. Satterthwaite. (1992). Sectoral growth patterns at the metropolitan level: an evaluation of economic development incentives. Journal of Urban Economics 31: 25-58.

O'Brien, R. (1992). Global Financial Integration: The End of Geography? London: Pinter.

O'Malley, E. and Vanegeraat, C. (2000). Industry Clusters and Irish Indigenous Manufacturing: Limits of the Porter View, Economic and Social Review, 31, 4, pp. 55-

OECD-DATAR 2001 World Congress on Local Clusters, Paris: OECD

Ohlin, B. (1933). Interregional and International Trade. Cambridge, MA: Harvard University Press.

Ohmae, K. (1995). The End of the Nation State: The Rise of Regional Economies, London: Harper Collins.

Pandit N. and G. Cook (1999). The Dynamics of Industrial Clustering in UK Financial Services. European Association for Research in Industrial Economics (EARlE), Turin, September 1999

Pandit, N. R. (1996). The creation of theory: A recent application of the grounded theory method. The Qualitative Report, 2(4): 1-(20.

Pandit, N. R., Cook, G. A. S. and Swann, G. M. P. (2001). The dynamics of industrial clustering in British financial services. The Service Industries

Journal, 21(4).

Pandit, N., G. Cook and G.M.P. Swann (1999). The Dynamics of Industrial Clustering in UK Financial Services", Working Paper no. 399, Manchester Business School, University of Manchester, June 1999

Panzar, J. C., and R. D. Willig. (1981). Economies of scope. Papers and Proceedings of the American Economic Association 71: 268-72.

Peneder, M. (1995). Cluster techniques as a method to analyze industrial competitiveness. International Advances in Economic Research 1 (3): 295-303.

Peneder, M., and K. Warta. (1997). Cluster analysis and cluster oriented policies in Austria. Paper presented at OECD Workshop on Cluster Analysis and Cluster Policies, Amsterdam, Netherlands, 9-10 October.

Perroux, F. (1950). Economic space: theory and applications. Quarterly Journal of Economics 64: 89-104.

Perroux, F. (1988). The pole of development's new place in a general theory of economic activity. In Regional Economic Development: Essays in Honour Francois Perroux, edited by B. Higgins and D. J. Savoie, 48-76. Boston: Unwin Hyman.

Petrakos, G. C. (1992). Urban concentration and agglomeration economies: re-examining the relationship. Urban Studies 29: 1219-30.

Petrakos, G., and J. Brada. (1989). Metropolitan concentration in developing countries. Kyklos 42: 557-78.

Pinch, S. and Henry, N. (1999). Paul Krugman's Geographical Economics, Industrial Clustering and the British Motor Sport Industry, Regional Studies, 33, 9, pp. 815-827.

Piore, M. J., and C. F. Sabel. (1984). The Second Industrial Divide. New York: Basic Books.

Polenske, K. R. (1997). Competition, Collaboration, and Cooperation: An Uneasy

Triangle in Networks of Firms and Regions. Multiregional Planning Research Project Working Paper. Cambridge, MA: MIT Department of Urban Studies and Planning.

Pollert, A. (1988). Dismantling flexibility. Capital and Class 34: 42-75.

Porter, M. and Ackerman, F.D. (2001). Regional Clusters of Innovation, Washington: Council on Competitiveness (www.compete.org).

Porter, M. E. (1995). The Competitive Advantage of the Inner City, Harvard Business Review, 74, pp.61-78.

Porter, M. E. (1998). On Competition, Harvard Business School Press.

Porter, M. E. (1996). Competitive advantage, agglomeration economies, and regional policy. International Regional Science Review 19: 85-90.

Porter, M. E. and Solvell, 0. (1998). The Role of Geography in the Process of Innovation and the Sustainable Competitive Advantage of Firms. In Chandler, A., Hagstrom, P. and Solvell, 0. (Eds) The Dynamic Firm: The Role of Technology, Strategy, Organizations and Regions, New York: Oxford University Press. pp. 440-457.

Pouder, R. and St John, C. H. (1996). Hot Spots and Blind Spots: Geographical Clusters of Firms and Innovation, Academy of Management Review, 21, 4, pp 1192-1125.

Powell W, (1990). Neither Market Nor Hierarchy: Network Forms of Organization. Research In Organizational Behavior, Vol. 12, pages 295-336

Pratten, C. (1991). The Competitiveness of Small Firms. Cambridge: Cambridge University Press. pp. 413-431.

Prevezer, M. (1997). The Dynamics of Industrial Clustering in Biotechnology, Small Business Economics, 9, pp. 255-271.

Pyke, F. And Sengenberger, W., Eds (1992). Industrial Districts and Local Economic Regeneration. Geneva: International Institute for Labour Studies.

Rabellotti, R. (1997). External Economies and Cooperation in Industrial Districts.

Houndmills: Macmillan Press Ltd.

Rabellotti, R. (1993). Industrial districts in Mexico--The case of the footwear industry. Small Enterprise Development 4: 26-36.

Rabellotti, R. (1995). Is there an 'industrial district model'? Footwear districts in Italy and Mexico compared. World Development 23: 29-41.

Redman, J. M. (1994). Understanding State Economies Through Industry Studies. Washington, DC: Council of Governors' Policy Advisors.

Rees, J. (1989). Regional development and policy. Progress in Human Geography 13: 576-88.

Regional Technology Strategies. (1997). Exports, Competitiveness, and Synergy in Appalachian Industry Clusters. Report to the Appalachian Regional Commission. Chapel Hill, NC.

Richardson, H. W. (1973). Regional Growth Theory. London: Macmillan.

Richardson, H. W. (1974). Empirical aspects of regional growth in the United States. Annals of Regional Science 8: 8-23.

Robinson, E. A. G. [1931] (1958). The Structure of Competitive Industry. Digswell Place: James Nesbit.

Rocca, C. A. (1970). Productivity in Brazilian manufacturing. In Brazil: Industrialization and Trade Policies. Cambridge: Oxford University Press.

Rodriguez-Pose, A. (2001). Local Production Systems and Economic Performance in France, Germany, Italy and the UK, in Crouch, C., Le Gales, P., Trogilia, C. and Voelzkow, H. (eds) Local Production System in Europe: Rise or Demise? Oxford: Oxford University Press, pp. 25-45.

Roelandt, T. J. A., and P. den Hertog. (1999). Cluster analysis and cluster-based policy making: the state of the art. In Cluster Analysis and Cluster-based Policy: New perspectives and Rationale in Innovation Policy, edited by T. Roelandt and P. den Hertog. Paris: Organisation for Economic Cooperation and Development.

Roelandt, T. J. A., P. den Hertog, J. van Sinderen, and B. Vollaard. (1997). Cluster Analysis and Cluster Policy in the Netherlands. OECD Workshop Position Paper on Cluster Analyses and Cluster-based Policies (Amsterdam, 10-11.10.97). Paris: OECD Industrial Cluster Focus Group.

Roepke H., D. Adams, and R. Wiseman. (1974). A new approach to the identification of industrial complexes using input-output data, Journal of Regional Science 14, 15-29.

Romano A., Passiante G. (1997). Innovation Territorial System as Learning Organization of Local System - Innovation Virtual System. The Regional Science Association 37th European Congress proceedings, Rome, August 1997

Romer, P. M. (1986). Increasing returns and long-run growth. Journal of Political Economy 94: 1002-37.

Romer, P., (1990). Endogenous technological change, Journal of Political Economy, 98

Rosenfeld, 5. (2001). Backing into Clusters: Retrofitting Public Policies, paper available at www.oecd.org

Rosenfeld, S. (1994). Making Mountains Out of Molehills or Go with the Flow: Industrial Clusters and Public Policy. Chapel Hill, NC: Regional Technology Strategies, Inc.

Rosenfeld, S. A. (1992). Competitive Manufacturing: New Strategies for Regional Development. New Brunswick, NJ: Center for Urban Policy Research.

Rouvinen, P., and P. Ylä-Antilla .(1997). A Few Notes on Finnish Cluster Studies. OECD Workshop Position Paper on Cluster Analyses and Cluster-based Policies (Amsterdam, 10-11.10.97). Paris: OECD Industrial Cluster Focus Group.

Russo, M. (1985). Technical change and the industrial district: The role of inter-firm relations in the growth and transformation of ceramic tile production in Italy.

Research Policy 14: 329-43.

Sabel. C. (1989). Flexible Specialisation and the Re-emergence of Regional Economies, in Hirst, P. and Zeitlin, J. (Eds) Reversing Industrial Decline: Industrial Structure and Policies in Britain and her Competitors, Oxford: Berg, pp 17-70.

Sayer, A. (1990). Post-Fordism in question. International Journal of Urban and Regional Research 13: 666-95.

Schmitz, H. (2000). Does Local Co-operation Matter? Evidence from Industrial Clusters in South Asia and Latin America. Oxford Development Studies, 28, 3, pp. 323-336.

Schmitz, H. (1995). Small shoemakers and Fordist giants: Tale of a supercluster. World Development 23: 9-28.

Schmitz, H. and Nadvi, K. (1999). Clustering and Industrialization: Introduction, World Development Vol. 27, No.9, pp. 1503-1514.

Schoenberger E. (1987). Technological and organizational change in automobile production: Spatial implications. Regional Studies 21, 199-214.

Schoenberger, E. (1988). From Fordism to flexible accumulation: Technology, competitive strategies, and international location. Environment and Planning D: Society and Space 6: 245-62.

Schulenburg, J.-M. & Wagner, J. (1991). Advertising, Innovation and Market Structure: A Comparison of the United States of America and the Federal Republic of Germany. In: (Acs, Z.J. and Audretsch, D.B., Eds.) Innovation and Technological Change, pp. 160-182. New York: Harvester Wheatsheaf.

Schweitzer, F. (1998). Modelling Migration and Economic Agglomeration with Active Brownian Particles. adv. complex systems 1, 11-37.

Scitovsky, T. (1954). Two concepts of external economies. Journal of Political Economy 62: 70-82.

Scott A., Storper M.J. (1993). The wealth of Regions: Market Forces and Policy

Imperatives in Local and Global Context. Harvard Business Review, March-April

Scott, A. J. (1992). The Role of Large Producers in Industrial Districts: A Case Study of High Technology Systems Houses in Southern California. Regional Studies 26, 265-275.

Scott, A. J. (1986). Industrial organization and location: Division of labor, the firm, and spatial process. Economic Geography 63: 215-31.

Scott, A. J. (1988). Metropolis: From the Division of Labor to Urban Form. Berkeley: University of California Press.

Scott, A. J. (1992). The collective order of flexible production agglomerations: Lessons for local economic development policy and strategic choice. Economic Geography 68: 219-33.

Scott, A. J., and A. S. Paul. (1990). Collective order and economic coordination in industrial agglomerations: The technopoles of Southern California. Environment and Planning C: Government and Policy 8: 179-93.

Scott, A. J., and E. C. Kwok (1989). Inter-firm subcontracting and locational agglomeration: A case study of the printed circuits industry in southern California. Regional Studies 23: 405-16.

Scott, A. J., and M. Storper. (1987). High technology industry and regional development: A theoretical critique and reconstruction. International Social Science Review 112: 215-32.

Scott, A.J. (1998). Regions and the World Economy, Oxford: Oxford University Press.

Scott, A.J. (Ed) (2001). Global City Regions: Trends, Theory and Policy, Oxford: Oxford University Press.

Scottish Enterprise (1998). The Clusters Approach: Powering Scotland’s Economy into the Twenty-first Century. Edinburgh: Scottish Enterprise.

Segal Quince Wicksteed (2001). Study of the Information Technology, Communications and Electronics Sectors, Cambridge: SQW Ltd.

Segal, D. (1976). Are there returns to scale in city size? Review of Economics & Statistics 58: 339-50.

Shefer, D. (1973). Localization economies in SMSAs: a production function analysis. Journal of Regional Science 13: 55-64.

Shohet S. (1998). Clustering and UK Biotechnology, in Swann, G.M.P., Prevezer, M. and Stout, D. (Eds) ((1998) The Dynamics of Industrial Clustering: International Comparisons in Computing and Biotechnology, Oxford: Oxford University Press, pp. 194-224.

Signorini L. F. (1994). The price of Prato, or measuring the industrial district effect, Papers in Regional Science 73, 369-92.

Slater D. (2004). Cluster Evolution, \transformation and the Steel Industry Supply Chain in N.E.England. Regional Studies. Vol.38.1 pp.55-66 Febr. 2004

Smith, D. F., and R. Florida. (1994). Agglomeration and industrial location: an econometric analysis of Japanese-affiliated manufacturing establishments in automotive-related industries. Journal of Urban Economics 36: 23-41.

Snyder M, deSouza Briggs Z. (2003). Communities of Practice: A New Tool for Government Managers. IBM Center for the Business of Government, November 2003

Solinas, G. (1982). Labour market segmentation and workers' careers: The case of the Italian knitwear industry. Cambridge Journal of Economics 6: 331-52.

Steiner, M. and Hartmann, C, (1998). Learning with Clusters: A Case Study from Upper Styria, European Research in Regional Science, 8, pp.

Steiner, M. and Hartmann, C, (2001). Looking for the Invisible: Material and Immaterial Dimensions of Clusters, Paper presented at the Regional Studies Association Annual Conference on 'Regionalising the Knowledge Economy', November 21, London.

Stenberg, L., and A-C Strandell. (1997). An Overview of Cluster-related Studies and

Policies in Sweden. OECD Workshop Position Paper on Cluster Analyses and Cluster-based Policies (Amsterdam, 10-11.10.97). Paris: OECD Industrial Cluster Focus Group.

Sternberg, E. (1996). The sectoral cluster in economic development policy: Lessons from Rochester and Buffalo, New York. Economic Development Quarterly 5 (4): 342-56.

Stigler, G. J. (1951). The division of labour is limited by the extent of the market. Journal of Political Economy 59: 185-93.

Storper, M., and B. Harrison. (1991). Flexibility, hierarchy and regional development: The changing structure of industrial production systems and their forms of governance in the 1990s. Research Policy 20: 407-22.

Stough, R. R. (1997). Merging quantitative and expert response data in setting regional economic development policy: Methodology and application. Paper presented at the 19th annual research conference of the Association for Public Policy Analysisand Management, Washington, DC. Copy available at policy.gmu.edu/cra/MSQA.html.

Streit, M. E. (1969). Spatial associations and economic linkages between industries. Journal of Regional Science 9: 177-88.

Streit, M. E. (1977). Agglomeration economies and industrial linkage: a reply. Journal of Regional Science 17: 129-30.

Sunley, P. (1992). Marshallian Industrial Districts: The Case of the Lancashire Cotton Industry in the Inter-War Years, Transactions of the Institute of British Geographers, 17, pp. 306-322.

Sveikauskas, L. (1975). The productivity of cities. Quarterly Journal of Economics 89: 393-413.

Sveikauskas, L., J. Gowdy, and M. Funk. (1988). Urban productivity: city size or industry size. Journal of Regional Science 28: 185-(202.

Sveikauskas, L., P. Townroe, and E. Hansen. (1985). Intraregional productivity differences in Sao Paulo state manufacturing plants. Weltwirtschaftliches Archiv 121: 721-40.

Swann G.M.P. (1997). Information and Communication Technologies and the Distribution of Economic Activity. Presentation to Prometeo Liberato: Employment and New Information and Communication Technologies, Fondazione Rosselli, Turin (February 1997)

Swann G.M.P. (1998). Networks and Clusters', presentation to ESRC /IMI conference, Learning Across Business Sectors, University of Warwick, (September 1998)

Swann G.M.P. (1998). The Dynamics of Industrial Clusters', seminar at City University, London (May 1998)

Swann G.M.P. (1997). The Internet and the Distribution of Economic Activity' presentation to Conference on Internet Economics, London Business School Regulation Unit (January 1997)

Swann G.M.P. (1999). "Clusters", Presentation to DTI/ESRC Seminar, DTI Conference Centre, June 1999

Swann G.M.P. (1999). Key Issues in Measuring Clusters. Presentation to DTI Seminar on Clusters, Chaired by Minister for Science, London, September 1999

Swann, G. M. P. (1998). Clusters in the US computing industry, in: G. M. P. Swann, M. Prevezer and D. Stout, eds., The dynamics of industrial clusters: international comparisons in computing and biotechnology (Oxford University Press, Oxford) 77-105.

Swann, G.M.P. (1998). Towards a Model of Clustering in High Technology Industries, in Swann, G.M.P., Prevezer, M. and Stout, D. (Eds) (1998) The Dynamics of Industrial Clustering: International Comparisons in Computing and Biotechnology, Oxford: Oxford University Press, pp. 52-76.

Swann, G.M.P. (1993). Clusters in High Technology Industries. The Business

Economist, 25(1), 27-36 (1993)

Swann, G.M.P. (1999). Innovation and the Size of Industrial Clusters in Europe. In A. Gambardella and F. Malerba (eds.) The Organisation of Economic Innovation in Europe, Cambridge University Press (1999)

Swann, G.M.P. (1996). Technology Evolution and the Rise and Fall of Industrial Clusters. Revue Internationale de Systemique, 10(3), 285-302 (1996)

Swann, G.M.P. (1999). The Internet and the Distribution of Economic Activity. In S. Macdonald, J. Nightingale and D. Allen (eds.) Information and Organisation: A Tribute to the Work of Don Lamberton, Elsevier, (1999)

Swann, G.M.P., M. Prevezer and D. Stout (eds.) (1998). The Dynamics of Industrial Clusters: International Comparisons in Computing and Biotechnology, Oxford University Press 1998

Tabuchi, T. (1986). Urban agglomeration, capital augmenting technology, and labor market equilibrium. Journal of Urban Economics 20: 211-28.

Tapscott D. (1996). The digital economy. New York. McGraw Hill

Taylor, M. J. (1973). Local linkage, external economies and the iron foundry industry of West Midlands and East Lancashire conurbations. Regional Studies 7: 387-400.

Teece D., (1988). Technological change and the nature of the firm. In Technical change and economic theory (ed. by G. Dosi et al), London, Pinter Publisher

Temple, P. (1998). Clusters and Competitiveness: A Policy Perspective, in Swann, G.M.P., Prevezer, M. and Stout, D. (Eds) (1998) The Dynamics of Industrial Clustering: International Comparisons in Computing and Biotechnology, Oxford: Oxford University Press, pp. 257-307.

Tremblay, G. D. (1993). Moving towards a value-added society: Quebec's new economic development strategy. Economic Development Review, Winter, 18-20.

Turner, A. (2001). The Competitiveness of Nations: Myths and Delusions, Chapter 1 in Just Capital: The Liberal Economy, London: Macmillan, pp. 23-50.

U.S. Department of Commerce. (1994). Benchmark Input-Output Accounts of the United States, 1987. U.S. Government Printing Office, Washington, DC.

Van der Laan, H. B. M. (1997). Everything you always wanted to know about clusters, but were afraid to ask. Opening address, OECD Workshop on Cluster Analysis and Cluster Policies, Amsterdam, Netherlands, 9-10 October.

Van Dijk, M. P. (1995). Flexible Specialisation, The New Competition and Industrial Districts. Small Business Economics 7, 15-27.

Van Gemert, J C. (2000). Text Mining Tools on the Internet - An Overview. ISIS Technical Report Series Vol. 23. Dept. of Computer Science, University of Amsterdam

Venables, A.J. (1996). Localization of Industry and Trade Performance, Oxford Review of Economic Policy, 12, 3, pp. 52- 60.

Verspagen, B. & Schoenmakers, W. (2000). The Spatial Dimension of Knowledge Spillovers in Europe: Evidence from Firm Patenting Data. mimeo, ECIS and MERIT.

Waits, M. J. (1995. Preparing for global competition through „industrial clusters." Spectrum, Summer, 34-7.

Watts D.J., (2005). A Simple model of fads and Cascading Failures. http://www.santafe.edu/research/publications/workingpapers/00-12-062.pdf

Wheaton, W., and H. Shishido. (1981). Urban concentration, agglomeration economies, and the level of economic development. Economic Development and Cultural Change 30: 17-30.

Wheeler, D., and A. Mody. (1992). International investment location decisions: the case of U.S. firms. Journal of International Economics 33: 57-76.

Williamson, O. (1985). The Economic Institutions of Capitalism. New York: The Free Press.

World Bank (2000). Electronic Conference on Clusters, World Bank.

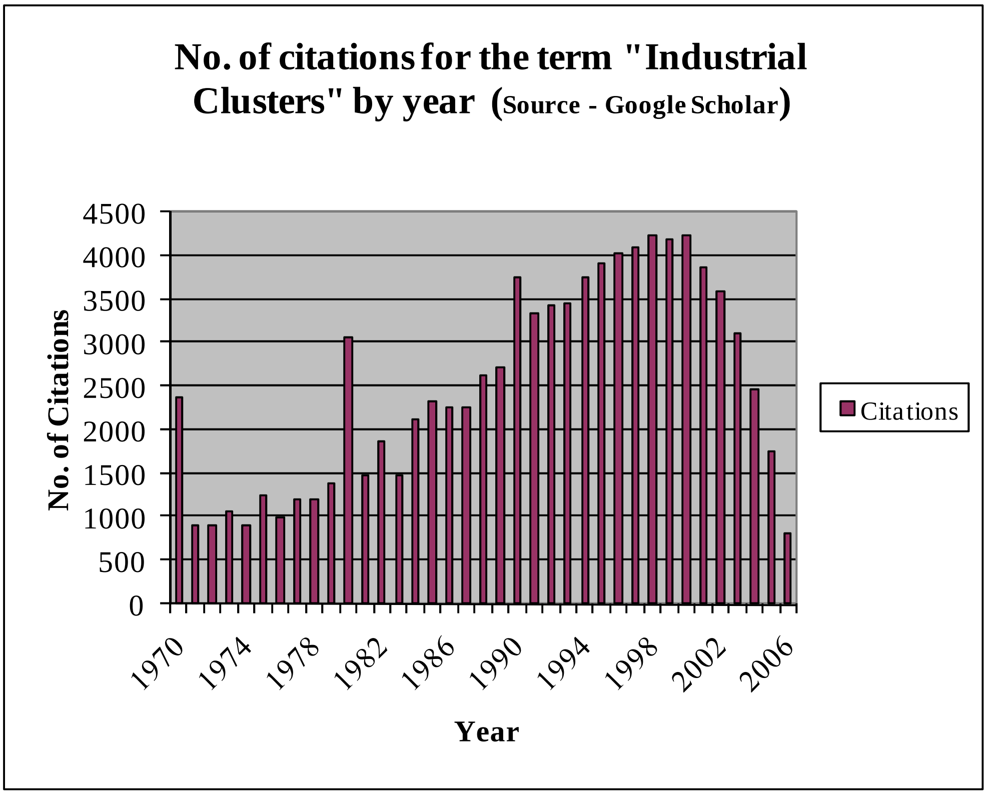
Young, A. (1928). Increasing returns and economic progress. Economic Journal 38: 527-42.

Zeisset, P. T., and M. E. Wallace. (1997). How NAICS will affect data users, Economic Planning and Coordination Division, U.S. Bureau of the Census, Washington, DC.

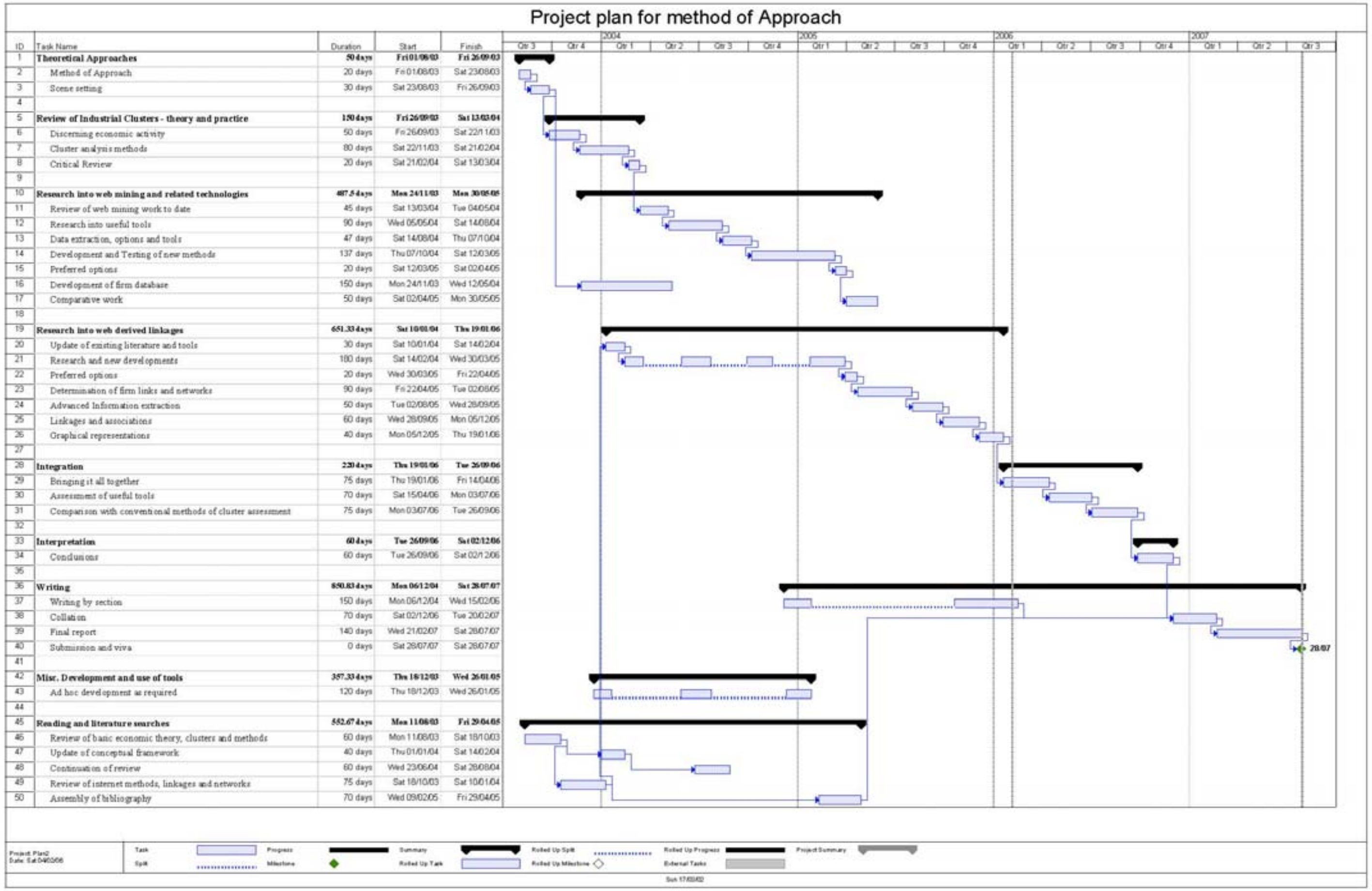
Zeitlin, J. (1995). Why are There no Industrial Districts in Britain? In Bagnasco, A. and Sabel, C. (Eds) Small and Medium Sized Enterprises, London: Pinter, pp.

Appendices

Appendix 1. Citations containing the clause "Industrial Clusters"- by year.



Appendix 2. Project Plan Outline



Project plan for method of Approach

ID	Task Name	Duration	Start	Finish
1	**Theoretical Approaches**	**50 days**	**Fri 01/08/03**	**Fri 26/09/03**
2	Method of Approach	20 days	Fri 01/08/03	Sat 23/08/03
3	Scene setting	30 days	Sat 23/08/03	Fri 26/09/03
4				
5	**Review of Industrial Clusters - theory and practice**	**150 days**	**Fri 26/09/03**	**Sat 13/03/04**
6	Discerning economic activity	50 days	Fri 26/09/03	Sat 22/11/03
7	Cluster analysis methods	80 days	Sat 22/11/03	Sat 21/02/04
8	Critical Review	20 days	Sat 21/02/04	Sat 13/03/04
9				
10	**Research into web mining and related technologies**	**487.5 days**	**Mon 24/11/03**	**Mon 30/05/05**
11	Review of web mining work to date	45 days	Sat 13/03/04	Tue 04/05/04
12	Research into useful tools	90 days	Wed 05/05/04	Sat 14/08/04
13	Data extraction, options and tools	47 days	Sat 14/08/04	Thu 07/10/04
14	Development and Testing of new methods	137 days	Thu 07/10/04	Sat 12/03/05
15	Preferred options	20 days	Sat 12/03/05	Sat 02/04/05
16	Development of firm database	150 days	Mon 24/11/03	Wed 12/05/04
17	Comparative work	50 days	Sat 02/04/05	Mon 30/05/05
18				
19	**Research into web derived linkages**	**651.33 days**	**Sat 10/01/04**	**Thu 19/01/06**
20	Update of existing literature and tools	30 days	Sat 10/01/04	Sat 14/02/04
21	Research and new developments	180 days	Sat 14/02/04	Wed 30/03/05
22	Preferred options	20 days	Wed 30/03/05	Fri 22/04/05
23	Determination of firm links and networks	90 days	Fri 22/04/05	Tue 02/08/05
24	Advanced Information extraction	50 days	Tue 02/08/05	Wed 28/09/05
25	Linkages and associations	60 days	Wed 28/09/05	Mon 05/12/05
26	Graphical representations	40 days	Mon 05/12/05	Thu 19/01/06
27				
28	**Integration**	**220 days**	**Thu 19/01/06**	**Tue 26/09/06**
29	Bringing it all together	75 days	Thu 19/01/06	Fri 14/04/06
30	Assessment of useful tools	70 days	Sat 15/04/06	Mon 03/07/06
31	Comparison with conventional methods of cluster assessment	75 days	Mon 03/07/06	Tue 26/09/06
32				
33	**Interpretation**	**60 days**	**Tue 26/09/06**	**Sat 02/12/06**
34	Conclusions	60 days	Tue 26/09/06	Sat 02/12/06
35				
36	**Writing**	**850.83 days**	**Mon 06/12/04**	**Sat 28/07/07**
37	Writing by section	150 days	Mon 06/12/04	Wed 15/02/06
38	Collation	70 days	Sat 02/12/06	Tue 20/02/07
39	Final report	140 days	Wed 21/02/07	Sat 28/07/07
40	Submission and viva	0 days	Sat 28/07/07	Sat 28/07/07
41				
42	**Misc. Development and use of tools**	**357.33 days**	**Thu 18/12/03**	**Wed 26/01/05**
43	Ad hoc development as required	120 days	Thu 18/12/03	Wed 26/01/05
44				
45	**Reading and literature searches**	**552.67 days**	**Mon 11/08/03**	**Fri 29/04/05**
46	Review of basic economic theory, clusters and methods	60 days	Mon 11/08/03	Sat 18/10/03
47	Update of conceptual framework	40 days	Thu 01/01/04	Sat 14/02/04
48	Continuation of review	60 days	Wed 23/06/04	Sat 28/08/04
49	Review of internet methods, linkages and networks	75 days	Sat 18/10/03	Sat 10/01/04
50	Assembly of bibliography	70 days	Wed 09/02/05	Fri 29/04/05

Project: Plan2
Date: Sat 04/02/06

Appendix 3. Clusters: A Variety of Definitions

(Some Examples Drawn from the Cluster Literature)

Porter (1998, 1999) "A cluster is a geographically proximate group of interconnected companies and associated institutions in a particular field, linked by commonalities and complementarities".

Crouch and Farrell, (2001, p. 163) "The more general concept of 'cluster' suggests something looser: a tendency for firms in similar types of business to locate close together, though without having a particularly important presence in an area".

Rosenfeld (1997, p. 4) "A cluster is very simply used to represent concentrations of firms that are able to produce synergy because of their geographical proximity and interdependence, even though their scale of employment may not be pronounced or prominent."

Feser (1998, p. 26) "Economic clusters are not just related and supporting industries and institutions, but rather related and supporting institutions that are more competitive by virtue of their relationships."

Swann and Prevezer (1996, p. 139) "Clusters are here defined as groups of firms within one industry based in one geographical area."

Swann and Prevezer (1998, p. 1) "A cluster means a large group of firms in related industries at a particular location".

Simmie and Sennett (1999a, p. 51) "We define an innovative cluster as a large number of interconnected industrial and/or service companies having a high degree of collaboration, typically through a supply chain, and operating under the same market conditions.

Roelandt and den Hertag (1999, p.9) "Clusters can be characterised as networks of producers of strongly interdependent firms (including specialised suppliers) linked to each other in a value-adding production chain".

Van den Berg, Braun and van Winden (2001, p. 187) "The popular term cluster is most closely related to this local or regional dimension of networks. Most definitions share the notion of clusters as localised networks of specialised organisations, whose production processes are closely linked through the exchange of goods, services and/or knowledge."

Enright (1996, p. 191) "A regional cluster is an industrial cluster in which member firms are in close proximity to each other."

Appendix 4 Website finder walkthrough

The reason for looking at this type of program was to gain a greater corpus of URLs than could be obtained from the sources as shown in Appendix 8. Chronology of URL database build. The basic idea recognises that in general terms lists of company names and addresses are easier to obtain than lists of URLs. Many agencies, as discussed in the main report sell such lists and if it were possible then to access a URL derived directly from a company's name and address it would clearly add to the general URL corpus. In some respects the process would seek to mimic a human user submitting a company name and address to a search engine. Having found some information regarding the company on say Google the next stage then takes the most appropriate top level domain name (URL) of the most promising looking result and looks up the details on the internet and compares the found company name and address details with the originally input company name and address. If this is a match i.e. the derived URL returns the correct company name and address as given on the original company list the found URL would be regarded as valid.

The simplified flow diagram in Figure 45 shows the basic elements of the URL finding process.

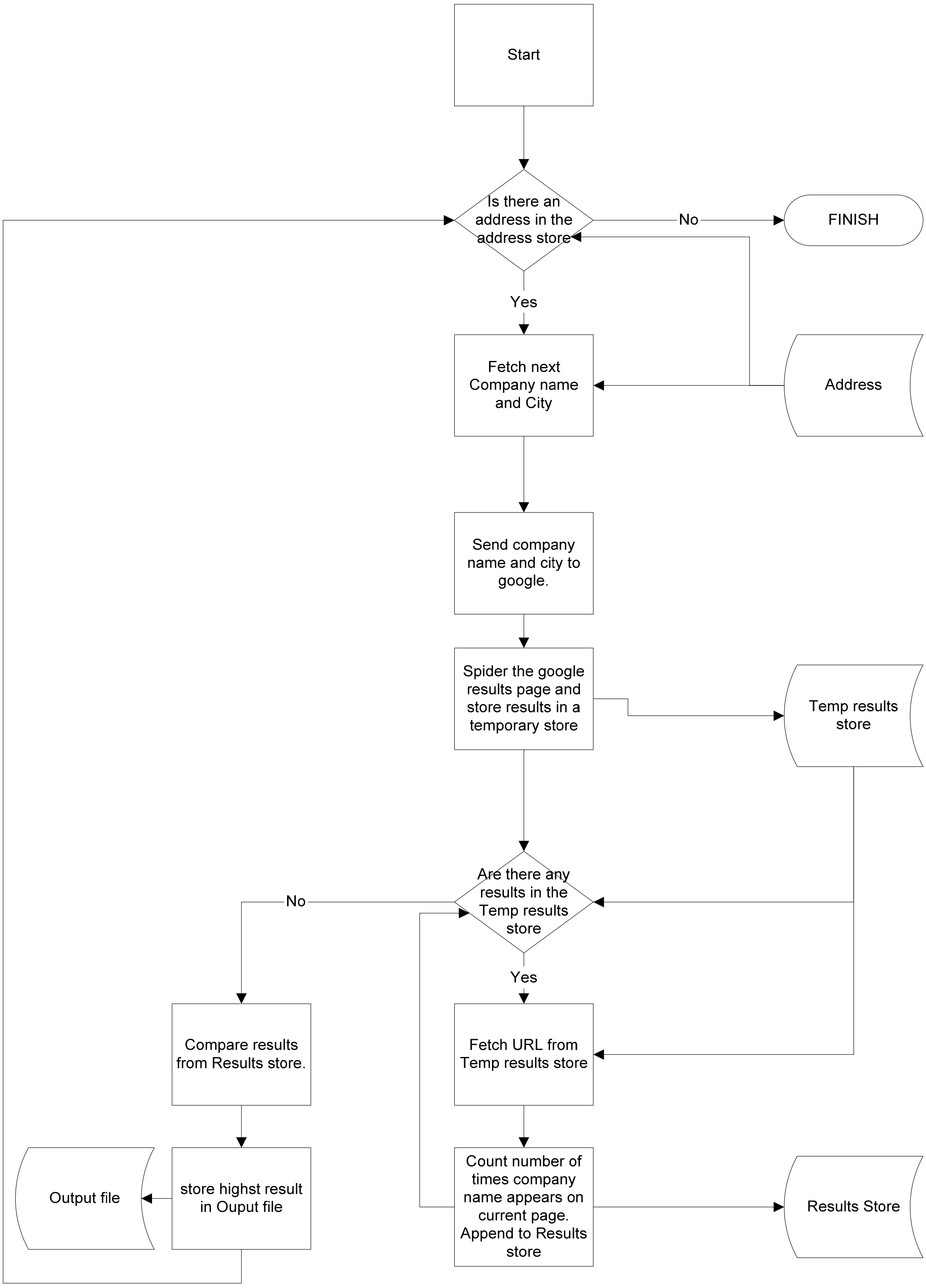
At the 'START' the program takes a company name and address details and then submits it to a search engine which returns the most probable URL of the website for the company. The term 'most probable' acknowledges that a perfect 'hit'on the company being sought is by no means certain.

The initial input of the program is a .csv (comma separated values) file of all the company names and addresses that the user is interested in. The program utilises the company name and the town fields. Provision was also made to use other fields apart from the town such as the postcode by placing the postcode in the town field within the csv file.   The company name and city are submitted to the UK version of Google with the company name as an exact phrase to be found, i.e. quotation marks round the company name viz "company name". The city is submitted with this as a normal submission being "company name" + city. All the links on the results page from Google are then placed in a temporary store.

The top level domain of the links is then extracted, for example "nemi-cai.co.uk/index.html" would become "nemi-cai.co.uk". This is to ensure that the page called is the homepage of the website. This extracted web address is then accessed and scanned by the program. Each time the company name is found the program increments a counter by one. Once the scan is complete the URL and the counter are then placed in a results area (if the company name is not found the counter will be recorded as zero). Once all the URL's have been completed the results area is opened and for each company the URL with the highest number is recorded to the output file. If a site is returned without finding the company name i.e. the counter is zero, then "No site found" is recorded in the output file.

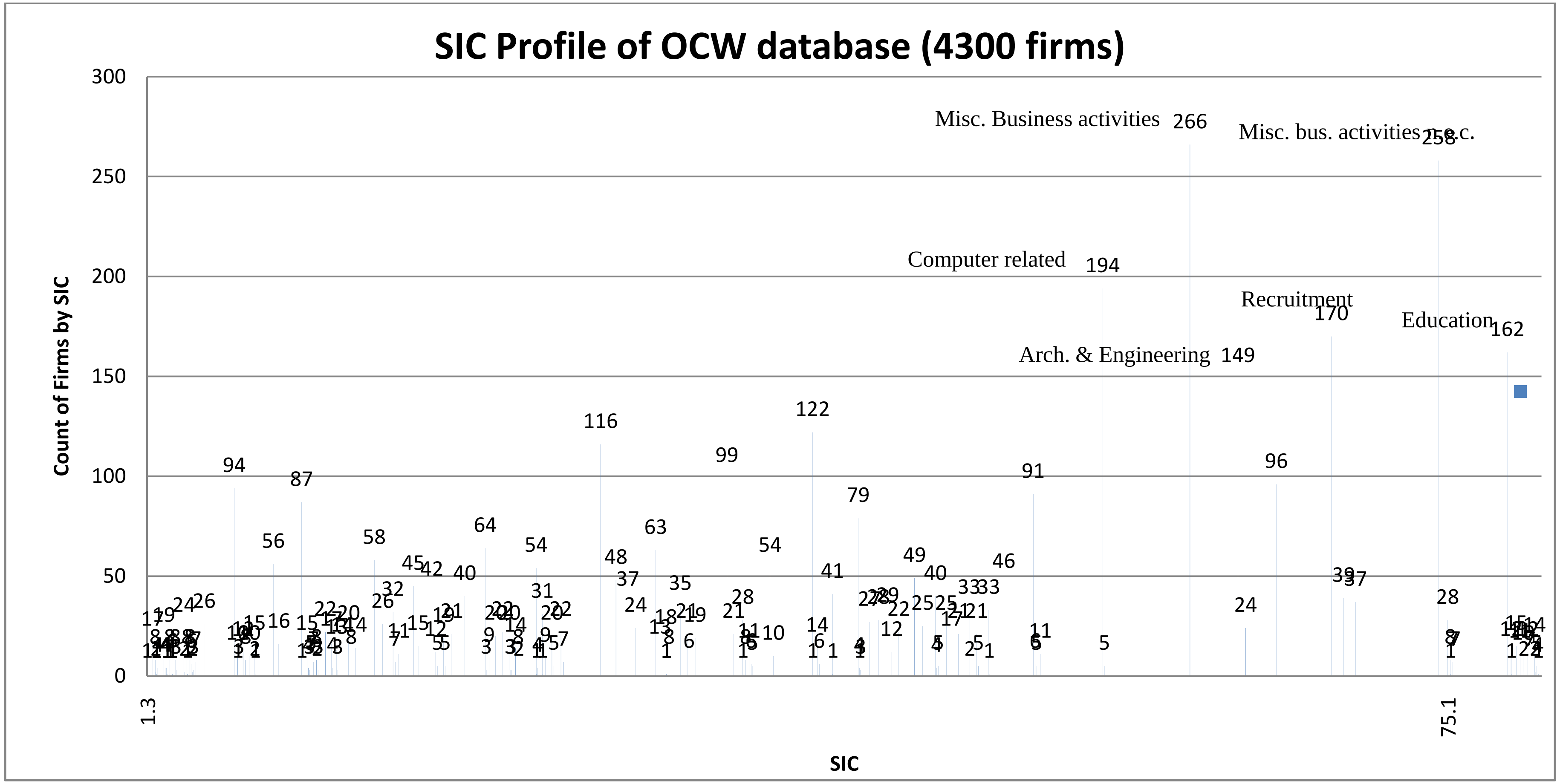
There are a number of practical issues not envisaged at the time of design of this program, caused largely by the presence of on-line trade directories. These are discussed in Section 8.3.2 Results and problems.

Figure 45 - URL finder - basic flow diagram


Start
Is there an address in the address store
No
FINISH
Yes
Fetch next Company name and City
Address
Send company name and city to google.
Spider the google results page and store results in a temporary store
Temp results store
Are there any results in the Temp results store
No
Yes
Compare results from Results store.
Fetch URL from Temp results store
Output file
store highst result in Ouput file
Count number of times company name appears on current page. Append to Results store
Results Store

Appendix 5 SIC Profiles

Figure 46 - SIC Profile of OCW Database



Notes.

Figure 46 - SIC Profile of OCW Database above shows the SIC make up of the 4300 firms in the OCW data. The predominant sectors are those associated with Business Services with Computer Related also represented. Architectural and Engineering also reflects the fact that the region has a substantial manufacturing base.

The top 10 by percentage are shown in Table 16 below.

Table 16 - Top 10 SIC by firm count

2 digit SIC	Count	% of total	Descriptor
74.1	266	6.2%	Misc. Business activities
74.8	258	6.0%	Business activities n.e.c.
72.6	194	4.5%	Computer related
74.5	170	4.0%	Recruitment
80.4	162	3.8%	Education
74.2	149	3.5%	Architectural & Engineering
55.1	122	2.8%	Hotels
45.2	116	2.7%	Building & Construction
51.5	99	2.3%	Wholesale
74.4	96	2.2%	Advertising etc.

The OCW data was assembled in 2003 with data that at the time could have been up to 2 years old. It is not known precisely how this would compare with the region overall as the data from each profile would have to be assembled on a like for like basis.

From what is known, by inspection it would seem that as would be expected business services are properly represented together with a manufacturing and construction base. If anything however Advertising, Hotels and possibly Computer Related might be over-represented as the region is not known as a particular hotbed of activity in these areas and again it might be more a reflection of the OCW raw data origins rather than actuality of occurrence.

Appendix 6 Communities of Practice - Definitions

[With reference to Section 2.5]

Bonding by exposure to common problems.

There are many shades of definition of this concept, but we define it as *"a group of professionals, informally bound to one another through exposure to a common class of problems, common pursuit of solutions, and thereby themselves embodying a store of knowledge."*

Peter & Trudy Johnson-Lenz, Awakening Technology [42]

Common practices and language

"Communities of Practice" is a phrase coined by researchers who studied the ways in which people naturally work and play together. In essence, communities of practice are groups of people who share similar goals and interests. In pursuit of these goals and interests, they employ common practices, work with the same tools and express themselves in a common language. Through such common activity, they come to hold similar beliefs and value systems.

Collaborative Visualization (CoVis) Project [43]

Commons sense of purpose

They are peers in the execution of 'real work'. What holds them together is a common sense of purpose and a real need to know what each other knows. There are many communities of practice within a single company, and most people belong to more than one of them.

John Seely Brown (2000)[44]

Groups that learn

Groups that learn, communities of practice, have special characteristics. They emerge of their own accord: Three, four, 20, maybe 30 people find themselves drawn to one another by a force that's both social and professional. They collaborate directly, use one another as sounding boards, teach each other.

[42] http://www.awakentech.com/

[43] http://www.covis.nwu.edu/info/philosophy/communities-of-practice.html

[44] http://www.fastcompany.com/magazine/01/people.html

Communities of practice are the shop floor of human capital, the place where the stuff gets made. Brook Manville, Director of Knowledge Management at McKinsey & Co., defines a community of practice thus: *"a group of people who are informally bound to one another by exposure to a common class of problem."* Most of us belong to more than one, and not just on the job: the management team; the engineers, some in your company and some not.

"The Invisible Key to Success" Fortune Magazine , August 5, 1996 by Thomas A. Stewart

Members evolve more creative practice

A community of practice is *"a diverse group of people engaged in real work over a significant period of time during which they build things, solve problems, learn and invent...in short, they evolve a practice that is highly skilled and highly creative."*

From a functional and experiential perspective, perhaps William Snyder and Xavier de Souza Briggs (2003 p.7) have provided the most practical definition to date for the 'community-of-practice' term in their research paper entitled **Communities of Practice: A New Tool For Government Managers.** Their view is that "*Communities of practice steward the knowledge assets of organisations and society".*

Appendix 7. Network Analysis Spider - Program Description

Outline

The processes that the Network Analysis Spider performs when given a set of URL's is shown in the accompanying flowchart in Figure 47.

The program starts with a given set of URL's provided by the user, which are saved to the BASE store. The spider checks if there is anything in the TODO store, first time round there will not be any so it carries on to check the BASE store.

It reads the first URL from the BASE store and scans the HTML.

All URL's are stripped from the HTML and sorted as either External or Internal URL's.

External URL's are stored in the EXTERNAL store with a reference to the BASE URL.

Internal URL's are stored in the TODO store.

Duplicate URL's are not entered into the either store.

The program then looks in the TODO store, if it finds a URL there it will follow that URL, read the HTML and store the URL's in the relevant stores.

Once the TODO store is empty OR the previously specified number of pages on a site has been reached, the TODO store is emptied and the next URL from the BASE store is read and the process starts again.

Once all the BASE URL's have been spidered, the results are generated.

By comparing each URL in the External store, with those in the original input file the spider produces a CSV file with the URLs on 2 axis grid. If a given URL has a link to a BASE URL then a 1 is printed at the intersecting point, if not then a 0 is printed.

Capabilities and Limitations

The spider will read any site that is written in HTML, PHP, ASP, JavaScript and most formats that serve text to the browser. The spider is limited by the fact that it cannot read Flash animated menu bars. There are a few sites that use flash to display the navigation of a given site. A similar type of problem was encountered with the keyword finding spiders in Sections 7.2 URL Scraper Program #8 and 7.4 'Phantom'.

As noted at the time HTML seeking programs cannot recognise Flash directly as Flash is basically a ‘picture’, a collection of pixels with no ‘meaning’. Although sophisticated pattern recognition programs do exist to discern words from graphics they are only partially successful, highly proprietary and unlikely to contribute anything meaningful in this context.

To try and combat this problem the spider pre-populates the TODO store for each site with various common names for external link pages such as “links.html”.

At a later stage in the programs development this was modified so that the spider always looked for the following pages first:

Index.asp?pageid=1
Default.aspx
Index.php
Default.asp
Index.asp
home.htm
default.htm
default.html
index.htm
index.html
links.html

These pages are read before the URL’s it finds on the first page. However, there is still a possibility that the site’s external links page does not get spidered.

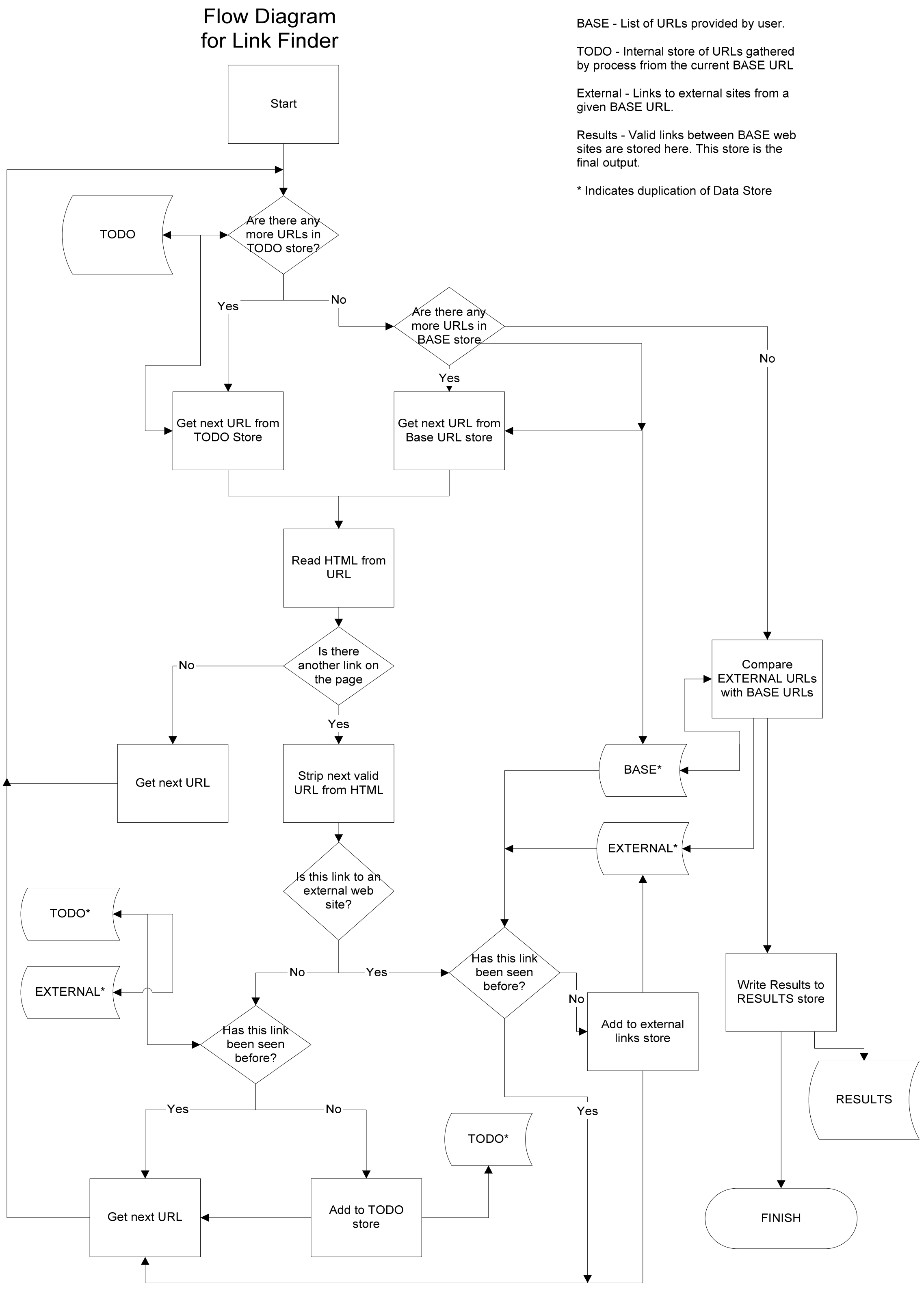
If the spider tries to fetch a page that either does not return any information or does not exist (dead link), it will try 5 more times then ignores that page and fetches the next URL from the TODO store.

Many websites redirect browsers from the front page to a different page. If the redirection takes the spider to another page under the same domain then the spider will follow the redirection and start from that page. If the redirected page takes the spider to a different domain, eg: Nemi.com to nemi-cai.com then the spider can not compare the results so will not show hits from that BASE URL.

The above is an outline of the basic schema but as more experience was gained with using the spider it was clear that several fixes would be required to deal with unusual domain names and internal redirects. In particular catalogue sites caused difficulties because of the very large number of components, each of which were spidered to the specified depth. This is analogous to problems found in Chapter 9. Further Work using the Combined Database regarding sites selling second hand cars in that the spider looked at and indexed all the details associated with thousands of individual cars up for sale on a site of this type.

These problems are discussed in the main body of the text in the context of testing and results.

Figure 47 - Network Analysis Spider – Program Flow Diagram


Flow Diagram
for Link Finder
BASE - List of URLs provided by user.
TODO - Internal store of URLs gathered by process friom the current BASE URL
External - Links to external sites from a given BASE URL.
Results - Valid links between BASE web sites are stored here. This store is the final output.
* Indicates duplication of Data Store
Start
TODO
Are there any more URLs in TODO store?
Yes
No
Are there any more URLs in BASE store
No
Yes
Get next URL from TODO Store
Get next URL from Base URL store
Read HTML from URL
Is there another link on the page
No
Yes
Get next URL
Strip next valid URL from HTML
Compare EXTERNAL URLs with BASE URLs
BASE*
EXTERNAL*
Is this link to an external web site?
TODO*
EXTERNAL*
No
Yes
Has this link been seen before?
No
Yes
Add to external links store
Has this link been seen before?
Yes
No
Write Results to RESULTS store
RESULTS
TODO*
Get next URL
Add to TODO store
FINISH

Appendix 8. Chronology of URL database build.

[Note: Parts of this appendix are used in context in the main body of the thesis where explanation of use is required]

The acquisition of company URLs has continued through the life of the project, partly by strategic intent and partly by opportunity. Originally the database of URLs was started with some 4300 as noted immediately below and the intention at that time was to use this corpus of regional URLs purely as a demonstrator for the methods being developed to add firstly useful information and then knowledge to the wider cluster debate particularly as regards discerning 'hidden' clusters that were not readily apparent from processing of SIC based company lists. There was no particular intention at the outset to gain a sample large enough for it to be strongly representative (in a statistical sense) of industrial activity in the North East as URLs on a large scale were not easy to acquire and in any case sectoral bias was deemed to be significant, though difficult to quantify (see Appendix 5 SIC Profiles). However as the research progressed a number of opportunities to acquire significantly more URLs arose from time to time driven in part by the general increase in use of the internet by companies and other organisations and also by various data collection agencies actually starting to log URLs as a field of interest. As a result of these opportunities further URLs were added, sometimes in small and very specialised groupings, at other times some authorities such as the North East Chamber of Commerce were, after some negotiation, able to hand over a significant number of company URLs derived from their membership lists.The net result of this process has been that the original database has more than tripled in size and as the data on which some of the research is based has become more robust a by-product has been a useful tool that has emerged for finding the capabilities of sampled firms within the region.
The timeline and background information for the various additional data sources follows below.

OneClickWizard (OCW)

The database of company URLs was originally started in 2003 with the acquisition of a set of 4300 obtained from OneClickWizard Ltd. This organisation was set up to acquire and market company data related to the North East of England and their business model was to give away the data on CD-ROM but have paid-for advertising from vendors which would be visible when the CD was used to search for goods and services. Although the business model in retrospect was somewhat flawed and soon overtaken by many on-line search engines doing a similar thing the company did spend time and effort checking the veracity of company details before committing them to the CD. This process was carried out by a small call centre located in the Centre for Advanced Industry, North Shields.

The original OCW database initially numbered over 13000 although this had been reduced to about 10800 validated names and addresses when the company ceased trading and of this number some 4300 firms had a URL that appeared to be valid from the response to a telephone call by OCW to the company (this is not quite the same as actually checking that the URL was a live link).

The provenance of the basic data i.e. how did OCW choose to acquire the basic company lists in the first place is a recurring theme with all URL databases that is discussed at various points in the main body of the report and also in this Appendix but in talking to the Principals of the firm the list acquired for checking by their call centre would seem to have been an amalgam of publicly available directories such as Thompsons, Yellow Pages and others such as Dun and Bradstreet lists. It was virtually impossible to know with any degree of accuracy how representative the final 13000 and hence the 4300 URLs were of regional industry. In addition in 2003 the 40% of these 13000 who did have a functioning website introduced a further potential source of bias in that some sections of business and industry such as IT Services and web designers were more likely to have a website than say small artisan companies. Although by observation this picture is changing over time (as would be expected as those who were not early adopters came on line with their own website) the only certainty of a true sample would be a near census as far as URL representation was concerned. This in itself would be a major project and outwith the scope of this thesis. It is perhaps worth noting in this context that publicly available lists are similarly incomplete. Trends Business Research using primarily longitudinal Dun and Bradstreet data in their company literature regards this database as a 'near census of companies over 10 people'[45]. However true this may be,

[45] Trends Business Research Ltd. Promotional literature 2006, www.tbr.co.uk

approximately 90% of firms employ less than 10 people so although D & B data may capture an economically significant proportion of the firm population, by definition it can hardly be said to capture a significant proportion of the total firms in a given area. Dun and Bradstreet data (now branded DNB) is discussed in context below.

As a final point many URLs appear on different data sources and it is important to remove duplicates. If this is not done then the same company is indexed more than once and indeed occasionally this does happen through oversight. Normally when new data sources were added they were checked against the existing combined database and any duplicates in the new data discarded.

However in the case of the DNB data (see below) the data were regarded as being firstly of high quality and more importantly as shown in Section 9.4 Comparative work it was felt desirable to preserve the DNB data as received and not remove any data that could be used for comparing information derived from the other DNB fields with the information gleaned from the RBKS. In the case of OneClickWizard records duplicates were therefore removed from this data if they appeared in the DNB data and not the other way around.

In the combined database the original 4300 OCW records when added to the combined database became 2136.

EMarket data

This data source is discussed in EMarket Data on page 132.

This data was acquired in February 2005 from a company called Mailing Tonic - www.mailingtonic.com based in Switzerland who supply for a few hundred pounds, company data including URLs under the product name ‘Emarket 2005’.

The perceived shortcomings and the decision not to use the data have been discussed in Section 8.4 Additional company records

In summary the initial attraction and enthusiasm associated with the prospect of 10004 records being available to the project soon evaporated when it became clear that virtually no quality checks had been applied to the data and the effort required to cleanse and sort the data into 4685 records even as a ‘first cut’ was out of all proportion to its usefulness.

As already noted if it was to be used as part of a wider dataset the downside could be the presence of a significant number of websites that added little to any cluster study and which could therefore distort the whole effort, a situation referred to in the early part of the discussion in Section 7.5 Full search of 4300 URLs using Phantom.

In spite of all the effort in cleansing these data every time a website was looked at manually a high proportion continued to show a dead link or some spurious or other entirely unsuitable website, at least from the point of view of industrial cluster research. This process of itself however did through up an interesting philosophical question and this was basically of the form:

'For any given population or associated set of URLs related to a region should any judgement be made regarding their suitability for inclusion in the study?'

Although the answer might seem obvious for a set of URLs such as those obtained from EMarket with its many apparently 'non-company' websites the answer is not in practice so clear. For example if one were to carry out a similar exercise based on Las Vegas where there is a significant sex industry it would be expected that there would be a great many sites that taken out of context, would be regarded as entirely unsuitable but to do so in a study of Las Vegas would be to discard a significant contributor to the local economy.

There is really no way to deal with this satisfactorily or consistently as a human is required to judge the suitability or otherwise of any particular site. This of course requires that the website is visited and scanned, at least for the top level. This process therefore largely obviates the whole idea of large scale information gathering mostly free from human intervention.

In view of these factors the decision was taken to discard all the work done with Emarket and look to other more robust data sources. The primary one that has been alluded to already was Dun and Bradstreet - DNB.

DNB.

In April 2005 arrangements were made with DNB to purchase a series of records in the four counties noted in DNB Data, these being Northumberland, Tyne and Wear, Durham and Cleveland. The count was 9002 records and of these some 8518 were found to be valid when submitted to the RBKS program. This is a relatively high proportion and was a welcome relief after the time consuming tribulations associated with EMarket.

Again although the quality of the URL data are good in that they refer to what appear to be highly relevant firms it is still a sample and only a sample. At the time of acquisition of the data there were about 42000 VAT registered firms in the four regions under discussion and even for 9002 firms that is still only 21.4% of the VAT registered total. It may be also that although firms that applied direct to DNB for a credit rating are also likely to appear on the VAT register, new and smaller firms picked off the Companies House register may be below the VAT limit and hence be not so registered. This proportion therefore if measured against all firms would likely be smaller still.

North East Chamber of Commerce (NECC)

The NECC in common with all Chambers is a business membership organisation. For the firm population size in the North East of England the NECC has a large membership and provides a comprehensive service of business support activities on behalf of its member companies.

In early May 2005 NEMI had been co-operating with the Chamber on an unrelated project and as part of a quid pro quo exchange of help and ideas NEMI reached an agreement for the Chamber to furnish NEMI with a list of Member URLs. A key requirement of the deal was that, in order to preserve a degree of anonymity regarding membership details only URLs would be disclosed and no other data would be provided.

For a conventional analysis of company activity using SICs or company activity descriptors such a URL database only would have been of very limited use but in the case of the RBKS the URL data was key. Although it would have been useful to have other fields in the manner of the DNB data for comparison purposes the 3648 new URLs added to the main URL database was a significant enhancement.

When duplicates with existing URLs were removed the new URLs added to the database totalled 1444.

Applegate Directory[46]

Applegate is an on-line directory focussed on firms in the Electronics, Engineering, Plastics, Chemical, Oil, Gas, Rubber and Recruitment Services in the UK. The site is easy to use for a small number of companies which can be searched by location,

[46] www.applegate.co.uk

type, or a number of other ways useful to market researchers. As Applegate were reluctant to sell their URL list even for North East England, in March 2006 a search related to electronics companies was undertaken by inspection. This was part of another project but did elicit 218 companies engaged in 'electronics' in some way and located within the four counties. However when duplicates had been removed the number of unique URLs fell to 68. What seemed to be happening at this point is that many of the better known or more substantive companies had already been picked up by one of the other databases and a lot of duplicates in the latest database became evident. As more databases are used as a source it would be expected that such effects would become more evident.

The North East Regional Portal (Tnerp)

The Regional Development Agency for North East England (One North East) undertakes a number of initiatives aimed *inter alia* at raising the economic prospects of the region. For a number of years it has been engaged in activities to promote the diffusion of broadband and the take up of internet and e-business facilities. As part of ONE's role in this area it formed a not-for-profit company called the North East Regional Portal (Tnerp) whose function was to channel initiatives in the use of e-business and the dissemination of information about the region. The access for the top level of the portal is www.n-e-life.com .

A separate function was to provide a website template and a small amount of financial assistance for small firms who did not have an existing web presence to acquire one. This is an ongoing process and firms become signed up with a website often in response to a marketing programme running at the time. Tnerp agreed to forward this list of URLs and this occurred over a period initially in April with further tranches to July 2006.

In all 4488 URLs were received from this source and of these 1989 were new i.e. not already on one of the databases. This is quite a high proportion (44.3%) as at this stage of the database build it would be expected that a much higher majority would have been accounted for already by the three main contributing databases being OCW, DNB and NECC. The figure can be compared with 39.6% being 'new' on the NECC database compared with what was already there.

The reason for this high proportion of new URLs is probably because many of these refer to small firms below the radar of data collection agencies and authorities such as DNB, VAT registrations and Companies House and also they are less likely to apply to any other credit rating agencies or be members of the Chamber.

Although it may add to the bias already alluded to, the firm information gained from the Tnerp data is an interesting addition and spidered information gained from the websites of these firms could not easily have been acquired by more conventional list based means.

Sub sea.

This was an entirely opportunistic acquisition as a result of work being done by Andriani and Siedlock (2006) on the emergence of subsea activity in the region. The names of the companies engaged in subsea activity found by a form of snowball sampling by the above authors were given. A surprising number, principally the smaller ones, did not appear to have a web site as when duplicates had been removed only 21 new firms of which 19 were valid were added to the combined URL database.

Other databases.

It is perhaps worth mentioning other well known databases and why they are not being used. Obvious ones would be Yellow Pages (Yell.com), Kellys and a variety of trade directories. There are two reasons; firstly as has already been mentioned this thesis does not seek to have a complete picture of North East regional industry via a near complete set of URLs. Secondly the data sources noted here are looked at periodically but in terms of their overall numbers the proportion of entries with a URL field completed is still surprisingly low. It seems that this is not that the firm does not necessarily have a website but more that the data collector involved has only in recent times started to log the URL as a field of interest. Another discouraging factor for firms promulgating their website is that some directories now also make an extra charge for a company URL being placed on the printed advertisement or line entry or for a clickable link on the directory website.

On the matter of the currency of the data although all the agencies claim to refresh their data on ‘regular basis’ without being specific about the actual timescale it is perhaps salutary that the rate of diffusion of URLs on to these databases still appears to be modest.

All of the se databases were incorporated in the RKBS starting sheet as shown below. As can be seen tick boxes allow the individual databases to be switched on or off in the search although for most practical purposes as the local search is so fast generally all the databases are used.

(c) J R Williams

2006

NEMI large scale keyword scan, 14000 starting URLs - North East England

Enter some key words to search by:

Find pages with any of these words and return 10 results.

Select session for search (select none for all sessions):

- □ 0-1000DNB (24/04/2005 - 717)
- □ 0-1000NECC (13/05/2005 - 894)
- □ 0-1000OCW (02/05/2005 - 784)
- □ 0-68applgelec (11/03/2006 - 61)
- □ 1000-1444NECC (14/05/2005 - 402)
- □ 1000-2000DNB (25/04/2005 - 838)
- □ 1000-2136OCW (02/05/2005 - 919)
- □ 2000-3000DNB (26/04/2005 - 811)
- □ 21Subsea (11/04/2006 - 19)
- □ 3000-4000DNB (26/04/2005 - 841)
- □ 4000-5000DNB (26/04/2005 - 819)
- □ 5000-6000DNB (27/04/2005 - 804)
- □ 6000-7000DNB (27/04/2005 - 828)
- □ 7000-8000DNB (28/04/2005 - 822)
- □ 8000-8518DNB (28/04/2005 - 427)
- □ NDI119 (28/02/2007 - 116)
- □ tnerp1 (10/04/2006 - 214)
- □ Tnerp2 (27/07/2006 - 595)
- □ Tnerp3 (27/07/2006 - 612)
- □ Tnerp4 (27/07/2006 - 568)

☑ Detailed Results □ Search Phonetically □ Begins With Searching

What's New: ○ Past Day ○ Past Week ○ Past Month □ Last Update

Search for key words found only in: □ URLs □ Titles □ Headers

□

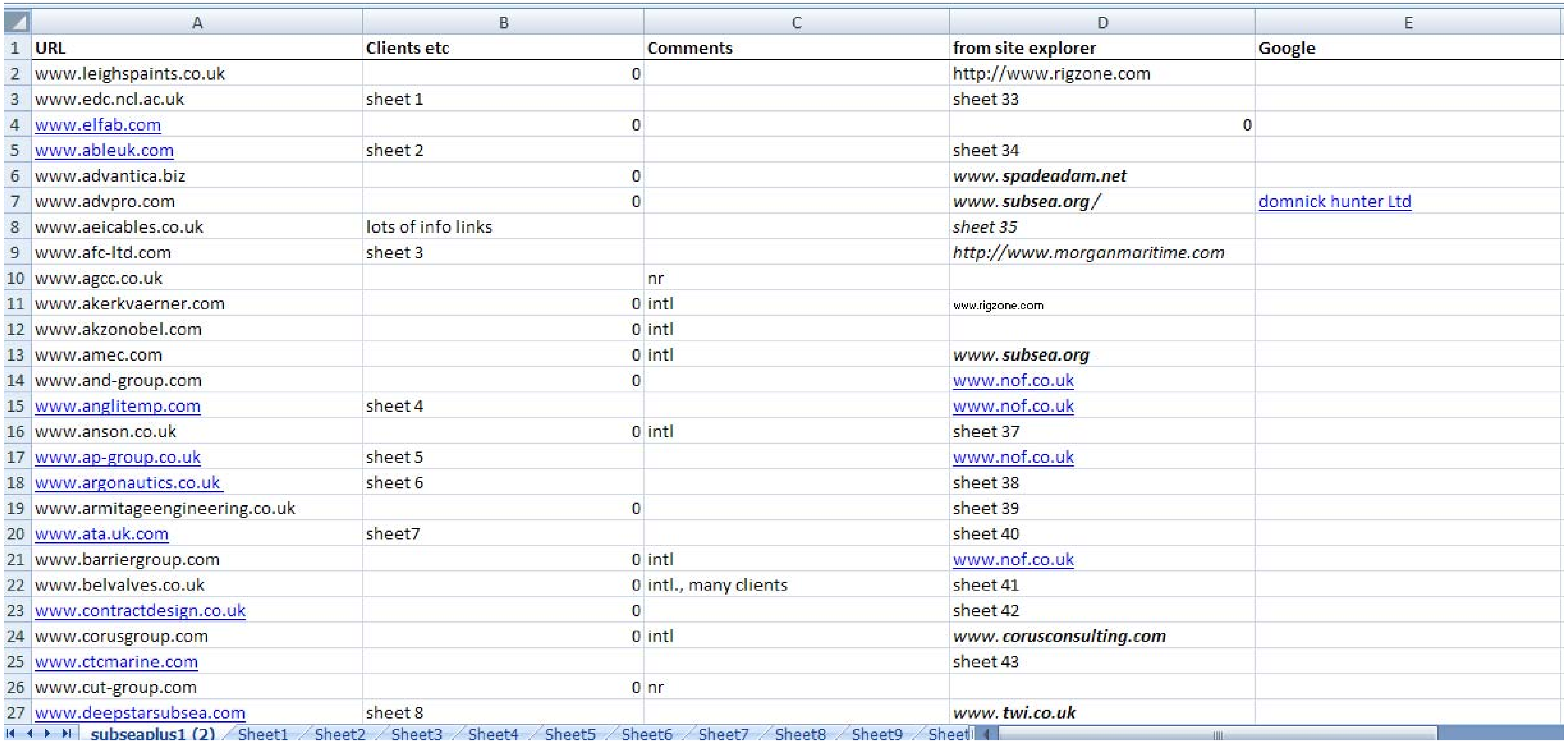
NEMI CAI
NORTH of ENGLAND MICROELECTRONICS INSTITUTE
CENTRE for ADVANCED INDUSTRY

jrw@nemi-cai.com / j.williams@ncl.ac.uk

Appendix 9. Example of Tracking Chart for In-links

Notes:

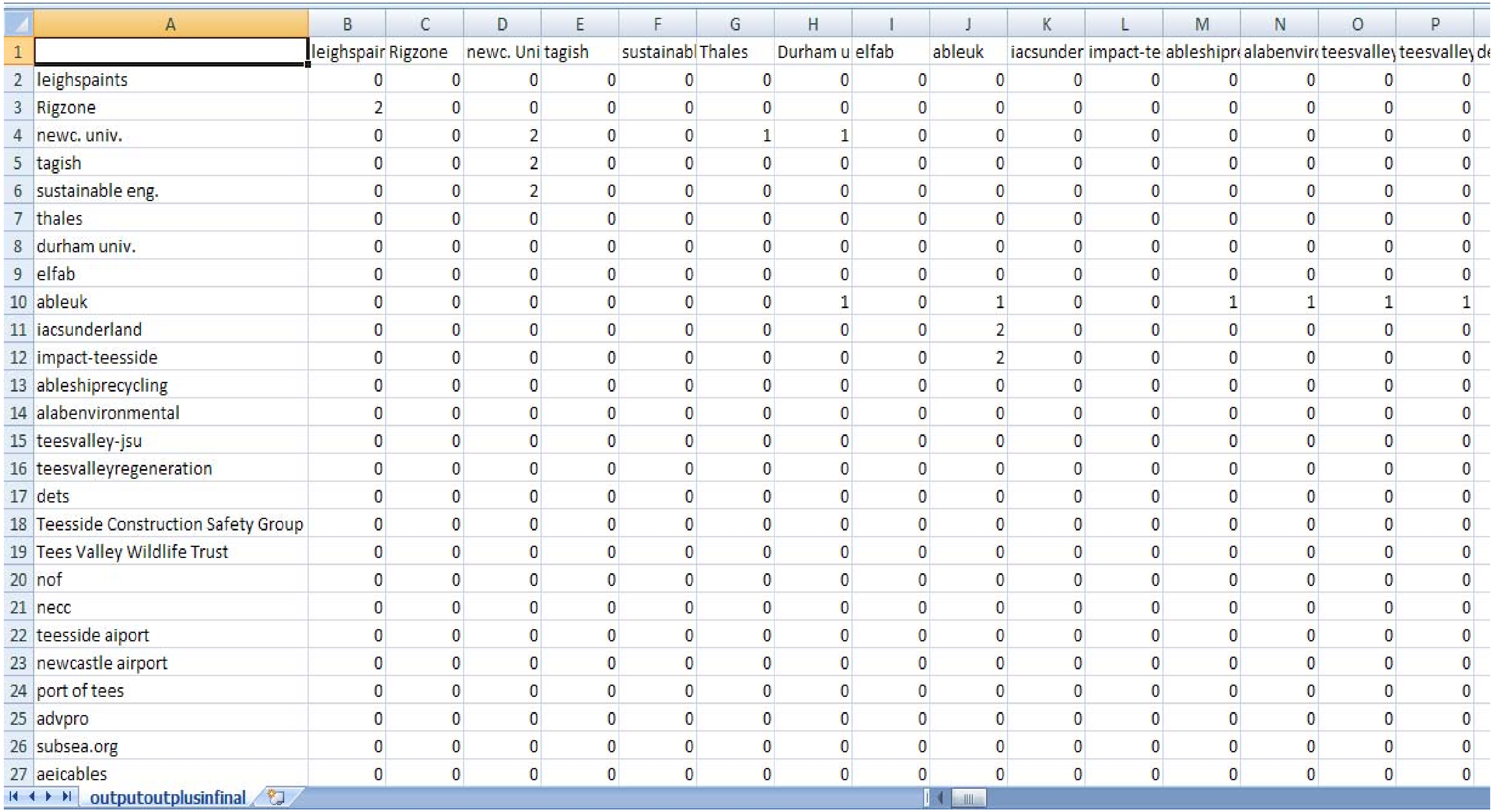
Column ‘D’ shows either the URL of a single in-link or the sheet number where the associated URLs are located if there is more than one. Column ‘B’ shows out-links in a similar manner. Any additional in-links from Google are in Column ‘E’.



	A	B	C	D	E
1	**URL**	**Clients etc**	**Comments**	**from site explorer**	**Google**
2	www.leighspaints.co.uk	0		http://www.rigzone.com	
3	www.edc.ncl.ac.uk	sheet 1		sheet 33	
4	www.elfab.com	0		0	
5	www.ableuk.com	sheet 2		sheet 34	
6	www.advantica.biz	0		*www.* ***spadeadam.net***	
7	www.advpro.com	0		*www.* ***subsea.org/***	domnick hunter Ltd
8	www.aeicables.co.uk	lots of info links		*sheet 35*	
9	www.afc-ltd.com	sheet 3		*http://www.morganmaritime.com*	
10	www.agcc.co.uk		nr		
11	www.akerkvaerner.com	0	intl	www.rigzone.com	
12	www.akzonobel.com	0	intl		
13	www.amec.com	0	intl	*www.* ***subsea.org***	
14	www.and-group.com	0		www.nof.co.uk	
15	www.anglitemp.com	sheet 4		www.nof.co.uk	
16	www.anson.co.uk	0	intl	sheet 37	
17	www.ap-group.co.uk	sheet 5		www.nof.co.uk	
18	www.argonautics.co.uk	sheet 6		sheet 38	
19	www.armitageengineering.co.uk	0		sheet 39	
20	www.ata.uk.com	sheet7		sheet 40	
21	www.barriergroup.com	0	intl	www.nof.co.uk	
22	www.belvalves.co.uk	0	intl., many clients	sheet 41	
23	www.contractdesign.co.uk	0		sheet 42	
24	www.corusgroup.com	0	intl	*www.* ***corusconsulting.com***	
25	www.ctcmarine.com			sheet 43	
26	www.cut-group.com	0	nr		
27	www.deepstarsubsea.com	sheet 8		*www.* ***twi.co.uk***	

Example of Chart used as input to graphing programs.

Notes: A cell with a ‘0’ signifies no connection, ‘1’ is an in-link and ‘2’ is an out-link. The sheet is a derived combination of multiple sheets as above that have either in-links or out-links.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	
1		leighspair	Rigzone	newc. Uni	tagish	sustainabl	Thales	Durham u	elfab	ableuk	iacsunder	impact-te	ableshipr	alabenvir	teesvalley	teesvalley	de
2	leighspaints	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	Rigzone	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	newc. univ.	0	0	2	0	0	1	1	0	0	0	0	0	0	0	0	
5	tagish	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	
6	sustainable eng.	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	
7	thales	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
8	durham univ.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
9	elfab	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
10	ableuk	0	0	0	0	0	0	1	0	1	0	0	1	1	1	1	
11	iacsunderland	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	
12	impact-teesside	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	
13	ableshiprecycling	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
14	alabenvironmental	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
15	teesvalley-jsu	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
16	teesvalleyregeneration	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
17	dets	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
18	Teesside Construction Safety Group	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
19	Tees Valley Wildlife Trust	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
20	nof	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
21	necc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
22	teesside aiport	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
23	newcastle airport	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
24	port of tees	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
25	advpro	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
26	subsea.org	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
27	aeicables	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Appendix 10. Examples of Output from OCW Tracking Cookies

Viewed Time	**Account Number Viewed**	**Account Number**
Bioprocessing Ltd:		
26/10/2001 14:03	1001241 mapo	1050691
26/10/2001 14:08	1001241 site	1049489
Castleton Communication Ltd:		
15/10/2001 07:08	1007 141 info	1049521
15/10/2001 07:09	1007141 site	1049521
15/10/2001 07:10	10071 41 site	1049177
15/10/2001 07:10	1007141 site	1049177
15/10/2001 07:10	10071 41 site	1049177
17/10/2001 13:51	10071 41 site	1050238
17/10/2001 09:35	10071 41 site	1050238
17/10/2001 09:35	1007141 site	1050238
17/10/2001 13:52	10071 41 mapo	1050238
17/10/2001 13:57	1007141 info	1042449
17/10/2001 13:57	10071 41 info	1050423
17/10/2001 13:57	10071 41 info	1050142
17/10/2001 13:58	1007141 info	1042449
17/10/2001 13:59	1007141 site	1042449
17/10/2001 14:00	1007141 site	1042449
17/10/2001 14:00	1007141 site	1042449
24/10/2001 09:47	1007141 info	1041125
24/10/2001 09:47	10071 41 info	1044959
24/10/01 09:47	10071 41 info	1033960
24/10/01 09:47	1007141 info	1048299
24/10/2001 09:47	1007141 info	1048399
24/10/01 09:47	1007141 info	1048299
24/10/2001 09:48	1007141 info	1041125
24/10/01 09:48	1007141 info	1048399
26/10/2001 14:11	1007 141 site	1050691
26/10/2001 08:41	1007 141 info	1049470
26/10/2001 08:42	1007141 mapo	1049470
26/10/2001 08:44	10071 41 info	1049549
26/10/2001 08:46	1007141 info	1050551
26/10/2001 08:48	10071 41 info	1046803
26/10/2001 08:49	10071 41 info	1042608

26/10/2001 08:49	10071 41 info	1043268
26/10/2001 08:49	1007141 info	1043440
26/10/2001 08:50	1007141 info	1042728

Identifier	**# hits**	**Name of Company**
1001241	2 Hits	Bio-processing Ltd
1007141	42 Hits	Castleton Communication Ltd
1013687	2 Hits	D & M Porter Ltd
1018963	6 Hits	EbacLtd
1021567	3 Hits	Executive Communication Service Ltd
1026020	41 Hits	Ambit New Media
1028244	12 Hits	Kall-Kwik Printing
1028406	269 Hits	K C Precision Engineering Ltd
1038084	15 Hits	Hadrian Air Conditioning Refrigeration
1041595	2 Hits	M D C Technology Ltd
1041727	6 Hits	Micro-chem Ltd
1041821	7 Hits	Murray Services Ltd
1042066	12 Hits	Olympus Products Ltd
1042455	6 Hits	Renpye Ltd
1042937	42 Hits	Taylor Packaging Ltd
1042961	1 Hits	Terra Nitrogen UK Ltd
1043276	3 Hits	U N N Commercial Enterprise
1043484	52 Hits	Pidra Enviroments
1043737	2 Hits	Raynor Whiting & Co
1045208	2 Hits	KTD Associates
1045609	4 Hits	North East Chamber Of Commerce
1047045	4 Hits	Bowe Computer Services
1047593	2 Hits	Formula Plastics
1047595	4 Hits	Foster Findlay Associates Ltd
1047740	142 Hits	One Click Wizard Ltd
1048181	2 Hits	MI Service Group
1048388	3 Hits	Oxley Developments Co
1048420	4 Hits	P C S Professional Computers

1049177	6 Hits	Bridge Recruitment
1049198	9 Hits	Central Employment Agency
1049290	16 Hits	Tyneside Training Services
1049343	5 Hits	Comfood
1049369	8 Hits	Dutton International Ltd
1049404	9 Hits	Entrust
1049467	13 Hits	Dev Engineering
1049470	252 Hits	Millenium Personnel Ltd
1049521	5 Hits	Temp Recruitment
1049564	27 Hits	Big Display Ltd
1049577	2 Hits	Regus Instant Office Worldwide
1049820	3 Hits	Milltech
1050078	2 Hits	Julia Smith Training Services
1050176	8 Hits	RTCNorthLtd
1050224	8 Hits	Bristol Street Motors
1050238	21 Hits	One North East
1050264	4 Hits	Avantium Technology UK Limited

Appendix 11 Social Networking in Business

The table below with accompanying notes are taken mainly from Pollard (Dec 2006) and are delineated by genre (or by 'taskonomy' i.e. what they are used for).

People-Connector Tools	**Examples**	**Useful for Identifying & Finding This Kind of People**	**What You Can Do Now**
People-Finders	LinkedIn, Ryze, Orkut, Facebook[1]	People meeting selected search criteria or having a specified affinity with you	Set up a just-in-time canvassing system[2]
Social Network Mappers	InFlow	People connected with others in an organization	Read The Hidden Power of Social Networks[3]
Proximity Locaters	DodgeBall	People you want to meet who are physically in your proximity	Use them to enable serendipitous meetings within your company[4]
Affinity Detectors	NTag (not free)	People with whom you have shared interests who are physically in your proximity	Use them at conferences where most attendees don't know each other[5]
Social Publishing & Info-Sharing Tools	**Examples**	**Useful for Publishing & Finding This Kind of Information**	**What You Can Do Now**
Journals	Blogs, Podcasts	Context-rich stories, reviews, and personal articles	Pilot blogs among those in the company already maintaining some sort of 'journal'[6]
Social Bookmarkers	Del.icio.us	Links to others' stories, reviews and articles (for those who don't have the time or interest to write their own blog)	Use del.icio.us to get standing notification of new articles on subjects of interest to your organization
Photo Journals	Flickr	Personal photos and visualizations	-
Memediggers	Digg, Reddit	Links to stories on 'hot' topics	-
Product Evaluators	Wize, ThisNext, Insider Pages	Consumers' evaluations of commercial products and services	Check out what potential customers are saying about the competition
Personal Diaries/ Music/ Video	MySpace	Information about and samples of people's favourite stuff	Put samples of your organization's possible new

Sharers			products on MySpace to test-market them
Collaboration and Communication Tools	**Examples**	**Useful for This Kind of Collaboration and Communication**	**What You Can Do Now**
Wikis	JotSpot	Simple, quick collaboration on document drafting and idea generation	Use wikis for small-group, ad hoc collaboration in your organization
Forums	Yahoo Groups	Threaded, subscribable conversations among communities of practice and communities of interest	Use forums for communication among ad hoc communities whose members are both inside and outside your organization
Commercial Collaboration Tools	BaseCamp (not free)	Project management including document sharing, discussions, scheduling, resource allocation, notifications	-
Mindmaps	Freemind	Real-time consensus-building in meetings and conferences; Visual representation of complicated information	Use mindmaps projected on a screen during meetings and conferences for instant documentation and resolution of misunderstandings
VoIP	Skype	Simple audio and video conferencing	Use Skype to enable free long-distance conferences when face-to-face is too expensive or impractical
Virtual Presence	Vyew	Real-time videoconferencing with screen-sharing, instant messaging, document sharing, whiteboarding, and attendance tracking	Use Vyew to enable small-group videoconferencing, virtual meetings, and training when face-to-face is too expensive or impractical
Peer Production	-	Producer-customer co-development of products and solutions (gift economy)	Read Umair Haque's paper and decide whether this technique has a place in your organization
'Unconferencing'	Open Space	Collaboratively addressing and resolving complex issues	Read Chris Corrigan's Open Space site and decide whether this technique has a place in your organization

Combinations of SNAs and Hardware	Mashups[7]		

Notes:

1. Facebook finds people within a specific school or organization.
2. Don't expect corporate directories to be current or give you the information you need to find true expertise. Instead, set up a just-in-time canvassing system, connected to e-mail groups around identified communities of practice in your organization, with request templates, to quickly find the people in your organization who have the expertise you need.
3. If you're going to map your organization's networks, use Rob's book to map the *value* of the networks, not just the volume of connections. Use it to support your just-in-time canvassing system (see above) and your communities of practice.
4. These tools avoid the embarrassment of rejection (and stalking) by notifying the person you are seeking to meet (rather than you) when the two of you independently indicate you are in close physical proximity; only when the other person responds positively to this notification are you notified that that person is willing and able to meet with you. This type of software has enormous potential to enable valuable meetings of people that would otherwise not occur.
5. These tools allow each tag recipient to key into the tag's 'smart' mag stripe information on their interests and expertise; when two people with shared interests and expertise come into close physical proximity, their tags 'flash' those shared interests and expertise to 'break the ice' quickly.
6. SMEs, CoP coordinators and internal newsletter editors are often ideal pilot groups for blogs, since they already have content that lends itself to journal format. Process: Identify the pilot group, select a blogging tool, develop and pre-populate a starting taxonomy, table of contents and initial content archive for each pilot member, develop appropriate security, RSS and internal/external access permission protocols, set up a help/monitoring group, offer everyone in the organization a brief seminar on blog publishing and subscribing, and talk up the externally-permissioned blogs outside the organization.
7. Examples: Health departments are using collaboration tools combined with Google Maps to map disease outbreaks; Caregivers are using wireless VoIP with GPS and digital monitors to allow seniors with medical conditions to live in their own homes and have their health monitored continuously and unobtrusively

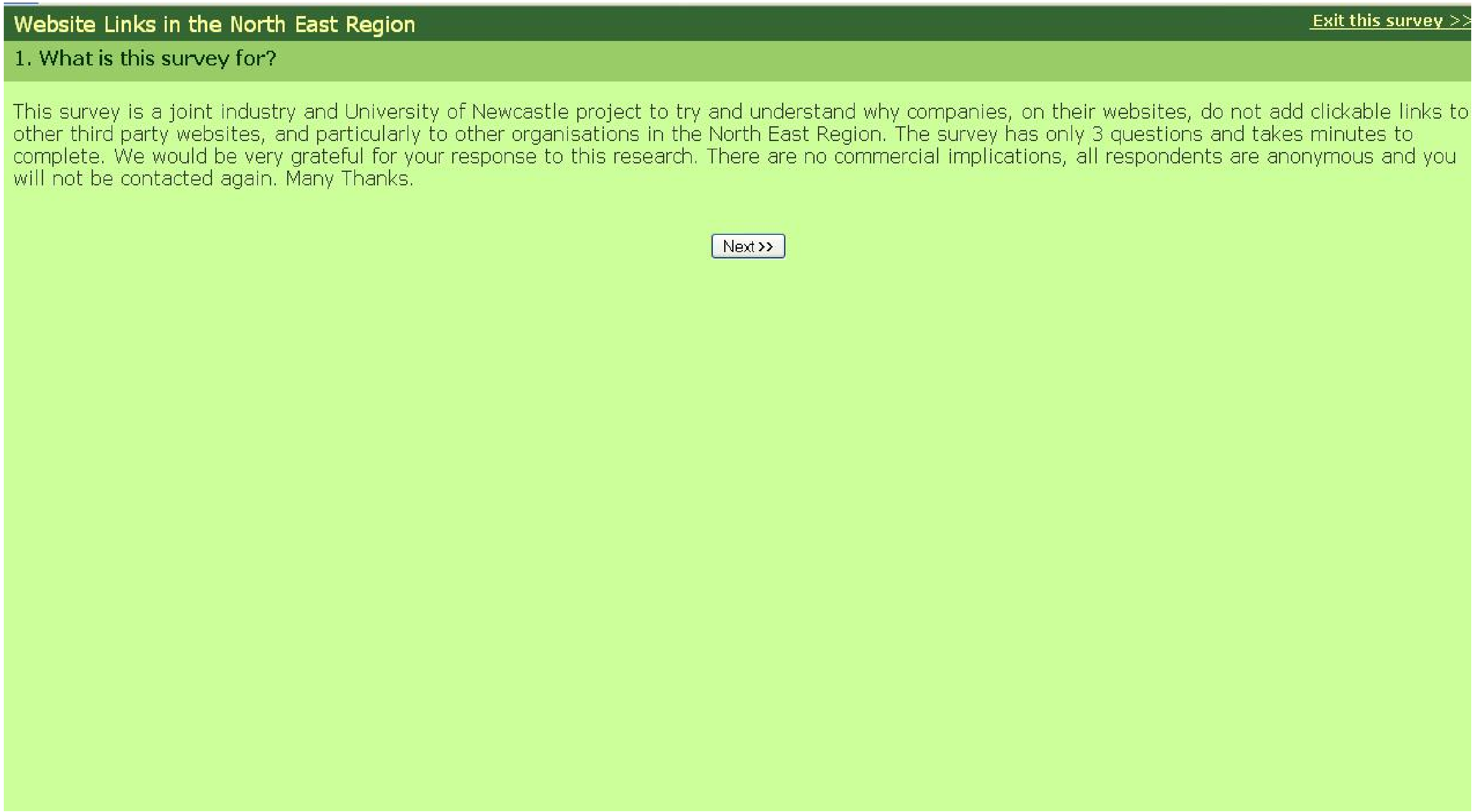
Appendix 12. Survey of Firms Web linking policies.

It was noted on page 201 that most firms did not think it worthwhile to embed hyperlinks to other firms or supporting organisations on their sites. This phenomenon was investigated with a simple emailed questionnaire to 200 companies who it was known from previous work e.g. architects had few if any clickable links on their websites. The response rate was in fact very low at c.5% which perhaps also indicates the level of interest in the subject. The two question sheets are shown below in Figure 48 and Figure 49 and the responses in Figure 50 and Figure 51.

Over half of responses to queries regarding why organisations did not put clickable links on their site seemed to be mostly related to their view that they could not see any advantages in doing so. This is mostly a problem of a lack of understanding of what the internet could do for the company and overt policies actively discouraging linking were in the minority.

In response to the question 'Do you plan to implement clickable links to more external sites in the future?', two thirds of respondents answered ‘yes’ or ‘possibly’ so the picture may change in the future.

Figure 48 - Survey - Introductory message



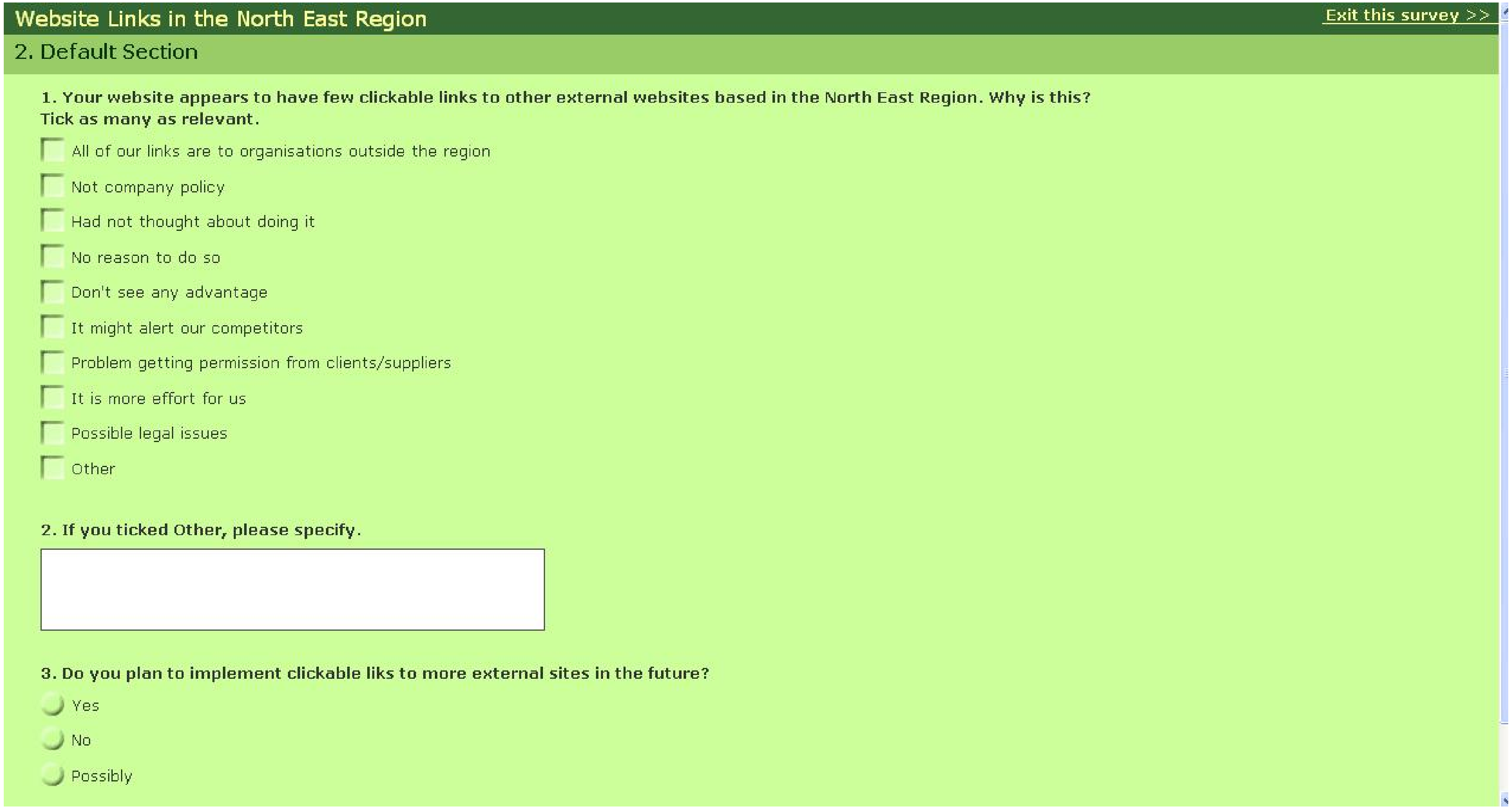
Website Links in the North East Region

1. What is this survey for?

This survey is a joint industry and University of Newcastle project to try and understand why companies, on their websites, do not add clickable links to other third party websites, and particularly to other organisations in the North East Region. The survey has only 3 questions and takes minutes to complete. We would be very grateful for your response to this research. There are no commercial implications, all respondents are anonymous and you will not be contacted again. Many Thanks.

The form for the user response is shown below in Figure 49

Figure 49 - User Response form



Website Links in the North East Region

2. Default Section

1. Your website appears to have few clickable links to other external websites based in the North East Region. Why is this? Tick as many as relevant.

□ All of our links are to organisations outside the region

□ Not company policy

□ Had not thought about doing it

□ No reason to do so

□ Don't see any advantage

□ It might alert our competitors

□ Problem getting permission from clients/suppliers

□ It is more effort for us

□ Possible legal issues

□ Other

2. If you ticked Other, please specify.

3. Do you plan to implement clickable liks to more external sites in the future?

○ Yes

○ No

○ Possibly

Figure 50 - Chart 1, user response

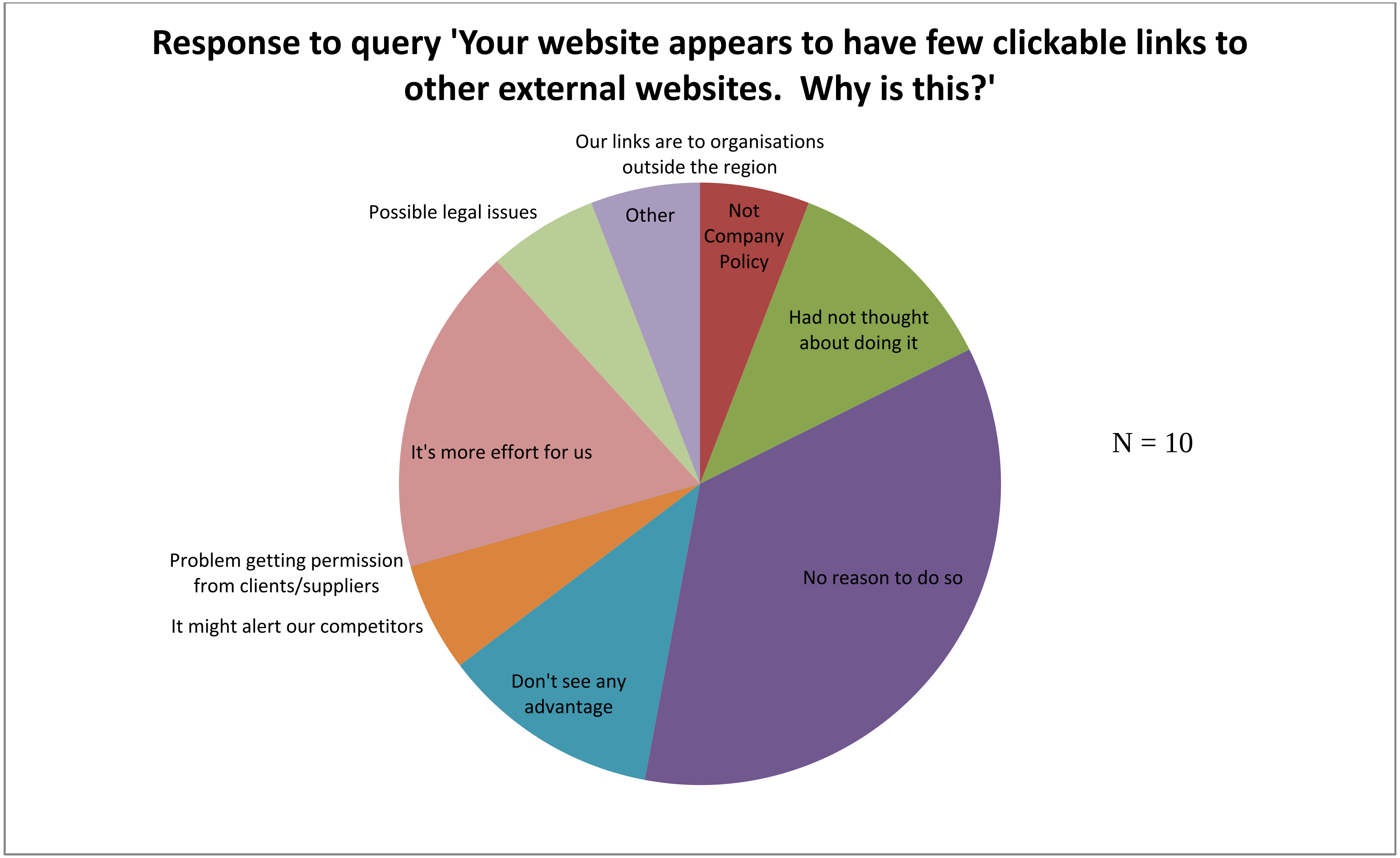
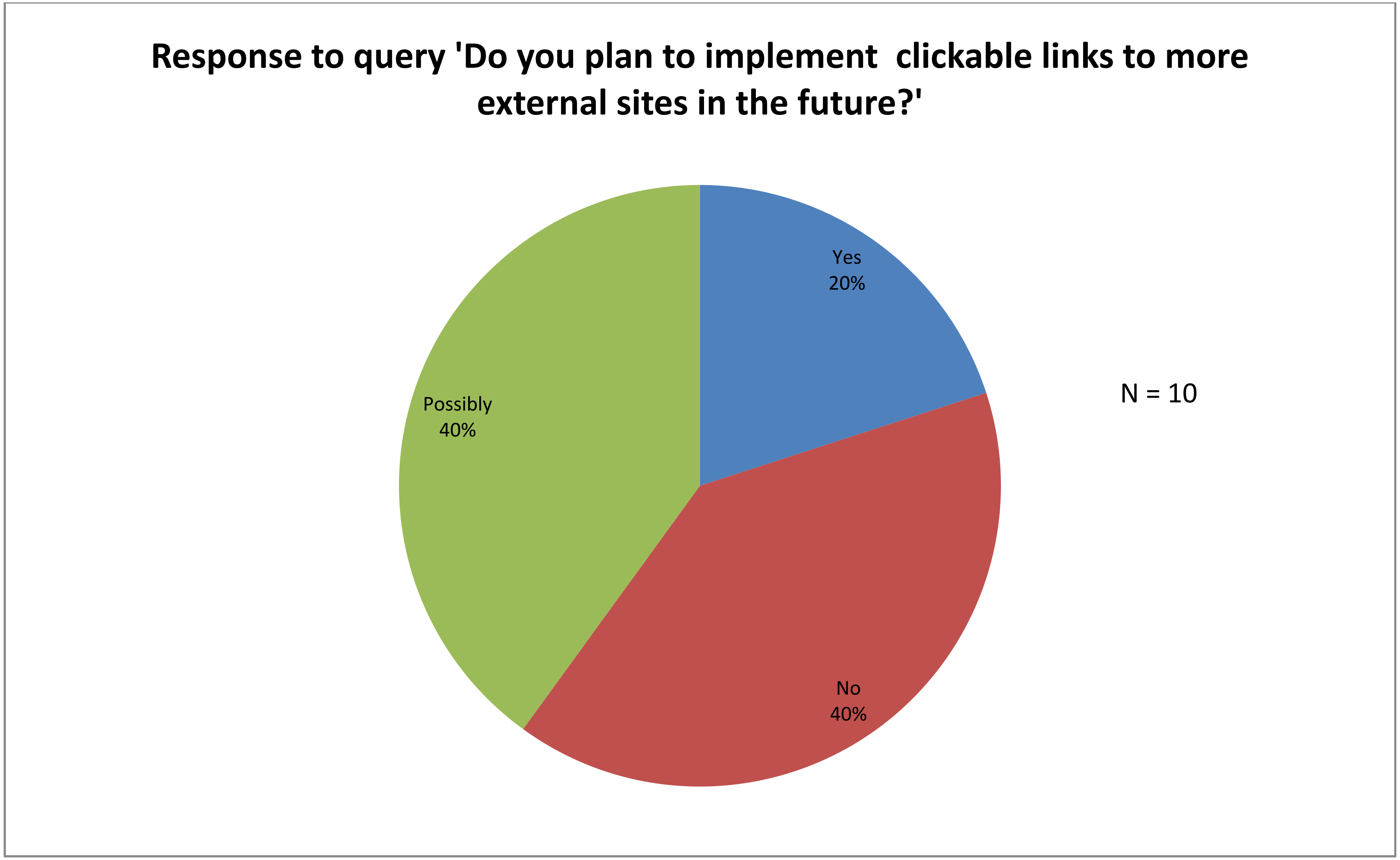


Figure 51 - Chart 2, User response



Response to query 'Do you plan to implement clickable links to more external sites in the future?'

Appendix 13 Definition of SNA terms.

These definitions are taken from Hanneman and Riddle (2005) and Wasserman and Faust (2004) insofar as they appear in the main text.

Betweenness centrality A measurement of centrality, indicating how 'powerful' an actor is in terms of controlling information flow in a network. The idea here is that an actor is central if s/he lies between other actors on the shortest paths connecting these actors.

Centralisation An index at group level, measuring how variable or heterogeneous the actor centralities are. It records the extent to which a single actor has high centrality and the other actors in the network have low centrality.

Centrality An index used to indicate how critical an actor is in a network.

Degree is the most popular way of measuring centrality. *See also* **betweenness centrality** and **closeness centrality.**

Clique A maximal complete subgroup in which all actors are directly connected to each other, and there are no other actors that are directly connected to all members of the clique.

Closeness centrality A measurement of centrality indicating how close an actor is to all other actors in a network. The idea here is that an actor is central if s/he can reach many other actors in a network in a few steps.

Clustering Co-efficient

Degree An index measured by the number of linkages adjacent to an Actor (node).

Density An index used to indicate how actors are closely or loosely connected in a network. It is measured by the proportion of possible linkages that are actually present in a network.

Distance based cohesion used to measure the ‘compactness of the network

Ego and Ego nets ‘Ego’ is an individual ‘focal’ node. A network has as many egos as it has nodes. A ‘neighbourhood’ is the collection of ego and all the nodes to whom ego has a connection at some path length. The nodes connected to the ego are sometimes referred to as ‘alters’.

E-I Index The (external – internal) index takes the number of ties of group members to outsiders , subtracts the number of ties to other group members and divides by the total number of ties. It is a measure of extent to which macro-structures ‘cluster’ the interaction patterns of individual nodes who fall within them

Eigenvector of Geodesic Distances The eigenvector approach is an effort to find the most central actors being those with the smallest ‘farness’ from others but in terms of the overall structure of the network.

Euclidean distance An index to measure structural similarity between actors of a network. The less two actors are structural equivalent, the larger the Euclidean distance between them. *See also* the Appendix.

Geodesic Distance For both directed and undirected data the geodesic distance is the number of relations in the shortest possible walk from one actor to another.

Hierarchical Clustering of Geodesic Distances. The geodesic distance is the length of the shortest path between nodes. A hierarchical clustering of distances produces a tree like diagram which helps to understand which nodes are most similar to one another and it can easily be portrayed for a small number of nodes.

Multi-dimensional scaling A way of visualising Euclidean distances. Networks are often visualised by two- or three-dimensional scaling in a graphical way with *(x, y,z)* coordinates, presenting a map of geometrical 'Euclidian' distances between actors in a network.

Overall graph clustering coefficient the average of the densities of the neighbourhoods of all the actors. A ‘weighted’ version gives weight to the neighbourhood densities proportional to their size.