Clustering of EPSRC research topics and researchers: using a network analysis approach based on grant data

by

Sergiu Tripon

sergiu.tripon.15@ucl.ac.uk

Supervisor:

Dr Shi Zhou, University College London

s.zhou@ucl.ac.uk

MSc Web Science & Big Data Analytics

Department of Computer Science

University College London

September 9, 2016

This report is submitted as part requirement for the MSc in Web Science & Big Data Analytics at University College London. It is substantially the result of my own work except where explicitly indicated in the text.

This side is purposely left blank.

Abstract

Collectively, humans take part in the everyday production of valuable data and intelligence with a significant use in areas including analysis, prediction and decision-making. The value of the data is primarily justified by the human judgement that contributed to its creation. Unfortunately, not everyone anticipates its value and the benefits it brings, and as a consequence, the opportunity of putting it to good use is often missed. Recently, EPSRC, an organisation which is in possession of substantial amounts of data, has been dealing with uncertainty with regards to defining research topics. Currently, it is unknown whether research topics should hold a more specific or broad definition. Additionally, once this is determined, how it could be achieved is also unknown. This models the problem that this research project aims to solve while also identifying an optimal solution in the process.

The primary objective of this research project is the application of a novel approach in graph theory to identify coherent clusters of topics within *Networks of Topics* constructed using current (2010 to 2016) and historical (1990 to 2000, 2000 to 2010) data collected from EPSRC. A secondary objective involves the discovery of researcher clusters through the analysis of *Researcher networks* using the same collected current and historical data.

A large-scale comparative analysis is carried out considering several network and edge weight interpretations and community detection algorithms with the aim of identifying an optimal solution which produces in the most well-defined, balanced, accurate and rational clustering of topics and researchers. The results show that the Louvain community detection algorithm applied on the *Topic (Grants as edges)* and *Researcher (Topics as edges)* networks using the normalised number of grants as the edge weight attribute resulted in the best topic and researcher clusters. This thesis proves that the novel approach to the problem is capable of making valuable use of the human judgement underlying the data.

This side is purposely left blank.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor, Dr Shi Zhou, for his kind assistance and constructive comments, as well as his indispensable guidance on the early direction of this thesis project.

Also, I wish to thank University College London for giving me the opportunity to study at such a reputable institution and for equipping me with the knowledge required to thrive during this year of study and further in life.

Finally, I would like to thank my amazing parents and girlfriend, for their endless support and encouragement throughout my study.

This side is purposely left blank.

Table of Contents

List of Figures

List of Tables

1 Introduction

Every day, humans produce valuable intelligence which can be used further in significant areas such as analysis, prediction, decision-making and many others. For example, when a user repeatedly watches TV series or films on Netflix, without realising, they provide Netflix with invaluable data in regards to their preferences including the genre, type and language. Their user account also provides useful information such as gender, age, when and how often they use Netflix, the device they access Netflix from and much more. This information can be analysed by Netflix as part of a research project in order to gather compelling insights from the data. The results of this research project can then have an impact on decision-making, for example, when deciding which series Netflix should add next to the platform based on user demand of a specific genre. Furthermore, it can also be used to improve the prediction of their recommender system. However, not everyone makes use of the gathered intelligence, and often, the opportunity to take advantage of it goes to waste.

One of the institutions that posses such important and substantial intelligence data is EPSRC. Currently, EPSRC holds a significant number of topics used by researchers to classify research grants when making a proposal. Furthermore, EPSRC are facing a difficult task in determining how finely or coarsely the topics should be defined. Subsequently, they are also uncertain on how this could be achieved, once the definition is determined.

This research project primarily seeks to cluster research topics into research areas using grant data collected from EPSRC. A secondary objective involves the clustering of researchers into researcher communities. The project aims to achieve this by applying a novel approach to a real world problem, involving graph theory, the concept of modularity, and community detection. Therefore, although the objectives are the clustering of topics and researchers, verifying whether this approach can be used to solve these

kinds of problems is also a crucial objective. It is expected that this approach will provide a cogent way to group topics in terms of similarity and aid decision making regarding their definition.

A number of different interpretations of the data collected from EPSRC were turned into Networks of *Topics* and *Researchers*. A further two different interpretations of each network were considered. Also, three different edge weight interpretations and eight community detection algorithms were considered. These amounts were refined to one optimal combination of edge weight interpretation and community detection algorithm, while the two different interpretations of each network were compared. Further details regarding the networks, edge weight interpretations and community detection algorithms are provided later in the thesis.

It was found that using the *Topic-grant* network produced a more rational topic clustering in comparison to the other interpretation of it, the *Topic-researcher* network. In contrast, the *Researcher-topic* network performed better than the *Researcher-grant* network when considering the researcher clustering. Furthermore, the edge weight, *weighted by the normalised number of grants* proved to be optimal. The results also showed that the *Louvain* method produced the most rational and well-defined clustering compared to the other community detection algorithms considered.

1.1 Structure of this Thesis

The remainder of this thesis is structured as follows. Chapter 2 provides an introduction to EPSRC and the problem addressed as well as concepts used throughout the project. Chapter 3 reviews the literature consulted during the project and highlights its contribution. The methodology followed throughout the project is described in Chapter 4. The results of the project are presented, compared and discussed in Chapter 5 and their evaluation is also documented. In Chapter 6, the work carried out throughout the project is summarised and concluded, and further potential work is recommended.

2 Background

In this chapter, background information regarding EPSRC is provided, while several concepts that are both related to and employed in the work carried out in this project, are also introduced.

2.1 EPSRC

EPSRC is one of seven research councils in the United Kingdom including *Economic and Social Research (ESRC)* and *Medical Research (MRC)* that form *Research Councils UK (RCUK)*, a non-departmental government body. RCUK's purpose is to manage the relationship between seven separate research councils which fund research projects in various disciplines.

EPSRC provides funding for research grants and training in two main disciplines: engineering and the physical sciences. Yearly, it invests more than £800 million in a variety of subjects ranging from mathematics to materials science, and from information technology to structural engineering [1]. It also boasts support worth £2-3 billion for a portfolio of research and training [2].

Currently, EPSRC holds a grant portfolio consisting of 3,175 grants. Each grant represents a research project within which one or more researchers collaborate, and it consists of substantial information including topic classifications. In total, the 3,175 grants are classified using 225 current topics. Researchers submit funding proposals to EPSRC, which, once accepted, turn into research grants financially supported by EPSRC.

2.2 Text-based methods

There are several types of clustering and classification methods, and most of them are text-based. This section introduces a number of them including *document clustering*, *topic Modelling* and *document classification*.

2.2.1 Topic Modelling

Topic modelling is a popular form of *text mining* used in *machine learning* and *natural language processing* to discover patterns in a text corpus. Naturally, a text body focusing on a particular topic such as computer games will contain certain words more or less frequently, *"graphics"* and *"animation"* more than *"shoes"* and *"dresses"*, for example. Other words such as *""is"* and *"the"*, also known as stop words, will appear frequently regardless of the topic, and are usually removed from the corpus due to their low value.

2.2.2 Document clustering

Document clustering is the task of dividing a document collection into a number of different clusters of documents based on similarity as a function of a document. It is a popular application in many areas such as information retrieval and topic extraction. There are clear differences between the classification and clustering of documents. The aim of document clustering algorithms is to divide a document collection into clusters of documents that hold a coherent structure. In contrast, classification is focused on discovering the type of a document by using its features.

Similarly to *topic modelling*, document clustering also analyses the text body of a document. However, there is a difference in motivation, as topic modelling is used to discover trends within text which could then be modelled into a number of topical keywords, representing the text as a whole.

2.2.3 Other forms of classification

Classification is the action of categorising a collection of entities into separate categories based on some criteria such as similarity. A *classification* represents an ordered list of the categories used to group the entities. Moreover, a

classification system is a method of realising *classification*. Furthermore, *classification* plays a valuable part in various subjects including *mathematics*, *media*, *science* and *business*. This section introduces a number of different forms of *classification* which share some common ground with the task that this project aims to accomplish.

2.2.3.1 Library classification

Library classification is the task of organising library material by subject or topic. Items are stored according to the order of the topics in the *classification*, which is represented by a *notational* system. This means that related materials are grouped in the same category, usually following a hierarchical tree structure.

In a library environment, the person responsible for classifying library materials is known as a *library cataloguer* or a *catalogue librarian*. The classification of a library material consists of two stages. Firstly, the cataloguer needs to find out what the material is about. This is followed by the material being assigned a call number by the *notational* system, which can be perceived as the address of a book. A library material can only be located in one physical space at a time, which means it can also only be assigned to one category at a time. In contrast, alphabetical indexing languages such as *Thesauri* or *Subject Headings* systems allow materials to be labelled with multiple terms.

2.2.3.2 Document Classification

Document classification is a *classification* problem in the fields of *library*, *information* and *computer* science. It deals with the task of assigning a document to one or more categories *manually* and *intellectually* or *algorithmically*.

Document classification is used in *library* classification where a *catalogue librarian manually* and *intellectually* determines what a library material is about, which results in its classification. In computer science, *document classification* is accomplished *computationally* through the use of various *document classification* algorithms.

2.3 Connectivity-based methods

In the previous section, a number of text-based clustering and classification methods were introduced. In contrast, this project does not aim to solve the problem defined using text-based methods, but connectivity-based methods such as network analysis, graph theory, the concept of modularity and community detection.

There are clear similarities and differences between the text and connectivity based methods. Firstly, this project holds contextual differences when compared to topic modelling and its applications. The former analyses a network of topics and aims to divide it into a number of coherent clusters while the latter is used to identify patterns as potential topics underlying a text corpus. Furthermore, *document clustering* and *community detection* have clear similarities in terms of the resulting rational clustering. In both methods, the members of a cluster hold a strong relationship. This is supported by weaker links and decreased similarity between members of different clusters. However, *document clustering* is based on the analysis of documents through a number of different techniques such as *tokenization*, *stemming*, *lemmatization*, removal of *stop words* and *punctuation*, and the computation of *term frequencies*. On the other hand, *community detection* is concerned with the structure of a network.

Subsequently, *library classification* holds concrete similarities with the motivation behind this project. EPSRC can be perceived as a substitute for a library, while grants are the equivalent of library materials. Both library materials and grants are classified using topics. Grants can be classified using one or more topics, while library materials must be assigned to a single category. Both forms of *classification* are performed manually and intellectually, one by a catalogue librarian and the other by researchers during the process of making a proposal for funding. However, the purpose of this project is not the classification of grants, but the grouping of research topics into different research areas. Moreover, the project aims to achieve this compu-

tationally, employing the human judgement underlying the data and a novel approach to the problem involving graph theory. In contrast, this is a step further than the library classification task where the process of classifying library materials is carried out manually by a human being.

Finally, *computational document classification* has similarities with the application of community detection as both are based on algorithms and achieved computationally. However, differences exist in terms of what is classified. Document classification aims to determine a document's category, while this project seeks to discover the research area representing a group of topics.

This section introduces a number of different network science sub-fields which are used throughout the project such as *network community structure*, *community detection* and the *concept of modularity*.

2.3.1 Network Community and Community Structure

A *network community* is a group of nodes which are densely-connected between each other and sparsely connected to other nodes in a network. Moreover, a *network community structure* is a group of network communities that is identified in a network.

Detecting the community structure in a network is one of several tasks in *network science*. Over the years, the task has become increasingly popular which led to the birth of a large number of community detection algorithms including *Spinglass*, *Louvain*, *Fast Greedy* and *Infomap*. A community detection algorithm divides a network into a number of clusters, which may be overlapping. If the nodes within each identified cluster are densely connected, it means that the network holds a community structure. In the case of no overlapping clusters, it means that the network is naturally divided and nodes within each cluster are densely connected while nodes between clusters are sparsely connected. It is assumed that two nodes are more likely to be connected if they share the same cluster, and less likely if they do not.

2.3.2 The concept of Modularity

Modularity, introduced by M.E.J. Newman [3], is one way to measure the strength of a community structure identified by a community detection algorithm. High modularity in networks means that nodes within each cluster are densely connected, while being sparsely connected to nodes in other clusters. Therefore, a network that holds a community structure is likely to also hold a high modularity. The motivation behind the concept comes from the analysis of social and biological networks which are known to hold a community structure. Identifying the community structure of such networks is invaluable in the journey of seeking a deeper understanding of a network's dynamics. For example, in a social network, information is more likely to travel faster within a community formed of densely connected nodes compared to a community composed of sparsely connected ones.

Initially, the modularity function solved the problem of dividing a network in two communities. The function was then modified so that it could also apply to the problem of network division into two or more communities. The concept of modularity is an extremely beneficial breakthrough in the network division problem.

However, it also has a resolution limit which leads to the inability of detecting small communities. During the process of dividing a network, a null model version of the network in question is created. A null model is an instance of a random graph which shares the same features as a specific real graph. The number of edges in a cluster within the real network is compared to the number of edges in a cluster within the null model. The null model assumes that each node can be linked to any other node in the network, which, if the network is large, is not necessarily true.

Regardless, the concept of modularity remains important and relevant as community detection algorithms often incorporate it in order to identify and measure the strength of a network's community structure.

3 Literature Survey

During the process of making a proposal, one of the tasks that researchers complete is the classification of the project using one or more research topics. By doing this, researchers use their judgement and provide intelligence in regards to the similarity between two grants based on topics and also the relationship between two topics based on grants.

A problem concerning the way topics within EPSRC are defined exists. It is not known how finely or coarsely the topics should be defined and how to define them in either way. This problem may arise when there are too many similar topics or not enough topics to cover the full topic classification spectrum of the grants. The relatedness of topics is also unknown. Currently, there are no known solutions or attempts to address the problem.

This project aims to identify a solution to the problem by using a novel approach involving graph theory and the concept of modularity. Focusing on grants that are classified by two or more topics, a network of topics can be constructed with the links between topics representing one or more common grants. Initially, within the network of topics constructed, there is no distinction between topics, apart from their names and links to other topics. Attempting to determine how topics should be defined while at the same time, focusing on the full network will prove extremely difficult. Subsequently, grouping similar topics manually, and analysing each group separately would lower the difficulty level, but will significantly increase the time it would take to complete the task.

In the end, the problem becomes a clustering problem, a very specific one, concerned with identifying the optimal way of dividing the network into clusters of topics in a rational and efficient manner. Furthermore, there are many other clustering problems in other fields such as *document clustering* which share some common ground with the problem addressed in this project. Furthermore, other fields are related to the approach used in this

project including *topic modelling* and *document classification*.

This chapter presents the latest development in a number of related topics, while it also introduces the state-of-the-art in *network analysis*, the main topic of this project.

3.1 Topic modelling

Topic Modelling represents a type of statistical model used to discover hidden topical structures within the text body of a document which is part of a document collection. It has become extremely popular over the years, with 17,043 research papers on Topic Modelling published since 2015, according to the ACM Digital Library [4].

In 2016, Gong et al. published *Who Will You @?* [5] which involves the development of a recommendation system focused on the mention function in microblogging services. In the process of recommendation, they system takes into consideration both the content of the microblog and the histories of candidate users. Moreover, a novel method which extends the translation-based model is proposed as a better way to handle the textual information. Experiments are carried out using a real-world data set collected from microblogging services. The results of the experiments prove that the proposed solution performed better than the previous state-of-the-art approaches.

In recent years, social media has cemented its popularity in the field of *topic modelling* as more and more researchers are attracted by the opportunity to analyse large volumes of text which are provided in data sets collected from social media services such as Facebook or Twitter.

This observation is justified by Sokolova et al. who, also in 2016, published *Topic Modelling and Event Identification from Twitter Textual Data* [6], which focuses on the analysis of four data sets of Twitter messages regarding challenging social events in Kenya. The *Latent Dirichlet Allocation (LDA)* model is used to analyse the text content, while the study is evaluated using both *Normalized Mutual Information (NMI)* and *topic coherence analysis*, in

order to identify the optimal LDA models. This study concludes that the tool developed has an effective use in the extraction of discussion topics for further manual analysis.

3.2 Document clustering

Document clustering is the process of applying clustering analysis to textual documents. Its use is divided between different fields such as topic extraction, fast information retrieval and document organisation. In contrast to *topic modelling*, the field of *document clustering* is not as popular, with 4,069 research papers listed in the ACM Digital Library [4], since 2015.

World Knowledge as Indirect Supervision for Document Clustering [7] was published by Wang et al. in 2016, and proposes a solution to the key obstacle that arises in making learning protocols realistic in applications, which is the need for supervision. Supervision represents a costly process as the involvement of domain experts is often required. The solution proposed represents a framework using world knowledge as indirect supervision. Furthermore, an example of using world knowledge for domain dependent document clustering is presented. Extensive experiments are carried out on several text benchmark data sets including *20newsgroups* and *Freebase*. The results of the experiments showed that incorporating world knowledge as indirect supervision is capable to outperform the state-of-the-art clustering algorithms as well as the ones enhanced with world knowledge features.

In contrast to the social-media focus identified in the latest *topic modelling* research, the research in the *document clustering* domain is much more varied. In 2016, Tripodi and Pellilo focused on *Document Clustering Games in Static and Dynamic Scenarios* [8], which proposes a game theoretic model for document clustering. In the game, each document to be clustered is a player and each cluster is a strategy. By interacting with others, players receive rewards. The game model is evaluated using 13 document collections using several different experiment settings. Compared to other document cluster-

ing algorithms, the results prove that the proposed solution performs well.

3.3 Document classification

In contrast to *document clustering*, *document classification* represents a problem in computer, information and library science that deals with the categorisation of a document into one or more categories.

However, similarities between the *document clustering* and *document classification* fields also exist in terms of popularity and research trend. Since 2015, according to the ACM Digital Library [4], 3,941 *document clustering*-related research papers have been published. Furthermore, research papers focusing on *document classification* are also diverse as certain publications focus on *classic document classification*, while others focus on specific areas of *document classification* such as *document image classification*.

The former is represented by *A Novel Approach to Document Classification using WordNet* [9], published in 2016 by Sarkar and Law. It aims to propose an alternative to the standard process used by other classification algorithms involving the bag-of-words approach to cluster analysis. The proposed alternative is based on *dictionary classification* and the correlation between words and phrases. The authors express their expectation of the solution proposed potentially leading to an improvement of the classifier's performance. However, it is not specified whether an improvement was actually achieved.

Document image classification, with a specific view on applications of patent images [10] by Gabriela Csurka, also published in 2016, involves a different type of *document classification*, as its main focus is *document image classification* and *retrieval*. Several different parameters for the *RunLength Histogram (RL)* and *Fisher Vector (FV)* based image representations are analysed and compared. Furthermore, an exhaustive experimental study is also carried out considering different document image data sets including the *MARG* benchmarks. The provision of guidelines on the optimal way of choosing the parameters in such a way that the features perform well in different tasks,

is the main aim of the study. The results of the experiment concluded that the suitable configurations for both features were suboptimal for individual tasks. However, in the situation where different tasks have to be solved with the same features, the proposed configurations are reasonable.

3.4 Network analysis

Network analysis is an academic sub-field of *network science* concerned with the study of complex networks such as social, biological and telecommunication networks.

Similarly to *topic modelling*, *network analysis* has become increasingly popular with time, especially since the rise of the *Internet* and *social media*, as researchers have developed a keen interest in the analysis of social networks. This translates in the number of *network analysis* research papers published since 2015, a total of 17,250 according to the ACM Digital Library [4].

On 3 December, 2014, a grand jury decided not to indict the white police officer involved in the death of Eric Garner. Motivated by this, in 2016, an extremely interesting study, *#Criming and #Alive: Network and content analysis of two sides of a story on twitter [11]*, was conducted by Kitzie and Ghosh. Furthermore, following the death of Eric Garner, the social networking platform Twitter was inundated with tweets sharing different opinions on racial profiling and police brutality. In order to analyse both sides of the story, the study compares tweets using two different hashtags: #CrimingWhileWhite (#cww) and #AliveWhileBlack (#awb). Furthermore, network and content analysis are employed on a large tweet data set containing the #awb and #cww hashtags. The study found clear differences in structure and the linguistic style between users, based on the used hashtag. Furthermore, it found that the #cww users shared informational content, while the #awb users were more subjective.

Moreover, *Network Volume Anomaly Detection and Identification in Large-scale Networks based on Online Time-structured Traffic Tensor Tracking* [12] pub-

lished in 2016 by Kasai et al., addresses a topological problem in the form of network anomography, the problem of inferring network-level anomalies from indirect link measurements. The study proposes an online subspace tracking of a *Hankelized* time-structured traffic tensor for normal flows. The abnormal flows are estimated as outlier sparse flows via sparsity maximisation. Furthermore, numerical-based experiments were carried out and results showed that the algorithm proposed achieves faster convergence and better volume anomaly detection performance when compared to the state-of-the-art algorithms.

3.5 Topic popularity in books

Out of pure curiosity, a search on the Google Books Ngram Viewer [13] was performed on the four research topics discussed above. The Google Books Ngram Viewer is an online search engine which calculates the frequencies of any comma-separated phrases found in materials printed between 1500 and 2008 within Google's text corpus. In this case, the time period selected was 1930 to 2008. Fig. 3.1 presents the result of the search. Surprisingly, the search engine was unable to find ngrams for *topic-modelling*.

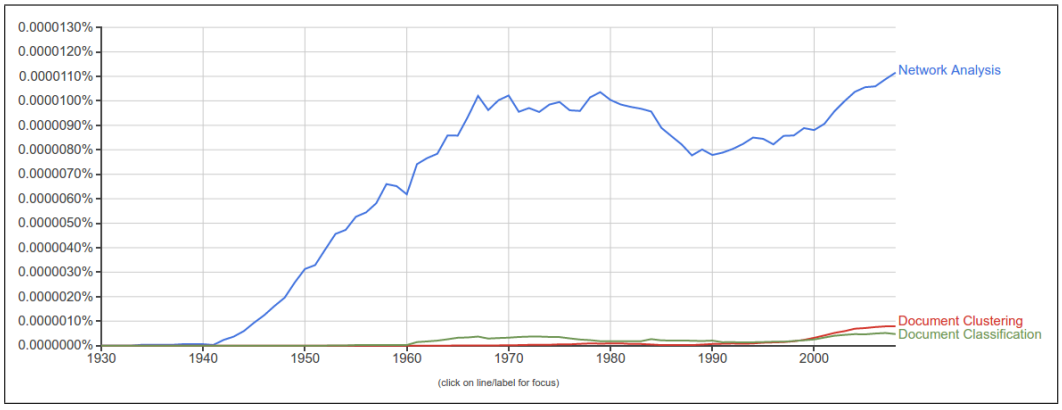


Figure 3.1: Frequencies of the four research topics discussed above found in materials printed between 1930 and 2008 within Google's text corpus. Surprisingly, the search engine was unable to find ngrams for *topic-modelling*.

4 Project Design

As a result of the project design phase, the problem statement and objectives were established, while the motivation behind the project was defined. A high-level sense of the methodology employed is also presented, together with a flow-chart created during the project planning phase.

4.1 Problem statement

The problem addressed in this project is concerned with the rational clustering of networks constructed using real data. This project focuses on identifying a solution to a real-world problem faced by EPSRC involving the definition of topic classifications. EPSRC is in possession of substantial grant data which can be used to formulate networks of *topics* and *researchers*. Ideally, the next step would involve the division of the networks created in order to obtain several groups of topics based on similarity or relatedness. The "best" group would represent the classification of its topics to a research area. This method would make the finer or coarser definition of topics significantly easier. However, as of yet, a way of achieving this has not been identified.

4.2 Objectives and Motivation

The overall aim of this project can be divided into three different objectives of diverse importance. Firstly, and most importantly, at the end of the project, a rational clustering of topics is expected to be produced using graph theory. Secondly, and equally as important, the project seeks to prove that a novel approach in graph theory can be employed to solve and provide a solution to the real-world problem addressed. Finally, this thesis project also aims to demonstrate that the novel approach can be applied to a different data set involved in the clustering of researchers.

Furthermore, the motivation behind this thesis project is primarily fuelled by the real-world aspect of the problem considered. Personally, I believe that contributing towards something that may be beneficial to other

people adds enthusiasm which translates into a more efficient work rate and higher standard of quality.

Originally, the concept of the project came from Dr. Shi Zhou, the supervisor of this thesis project. Recently, he attended a talk where an EPSRC officer expressed concerns regarding the way research topics were defined. Therefore, the project was principally specific to one organisation and problem. However, with time, the data took on a new significance as the invaluable human judgement underlying it was discovered. This changed the aim of the project as it became a project concerned with a general real-world problem. In this new meaning, graph theory combined with the human judgement behind the links connecting the topics in a network is used to identify an optimal way of clustering topics meaningfully.

4.3 Methodology

This project uses current and historical data publicly provided by EPSRC through the EPSRC Grants on the Web (GoW) service. Networks of *Topics* and *Researchers* are constructed using the data collected from EPSRC. Node and edge attributes are also formulated, and the attribute values normalised. This is followed by extensive comparison experiments on two different interpretations of the two types of networks considering three different edge weight interpretations and eight different community detection algorithms. The three edge weight interpretations are *unweighted*, *weighted by the normalised number of grants* and *weighted by the normalised value of grants*, while community detection algorithms include *Spinglass*, *Louvain* and *Fast Greedy*. Subsequently, the topic and researcher clusters are evaluated by calculating the average Dice and Jaccard similarity of node pairs between and within the clusters. The results of the experiments are expected to represent the identification of an optimal combination of edge weight and community detection algorithm which produces a coherent, balanced and well-defined clustering of topics and researchers.

4.4 Project Plan

Prior to development and analysis stages, a plan was produced in the form of a flow chart, presented in Fig. 4.1, which details every major process that was completed in order to achieve the objectives set in the Project Design.

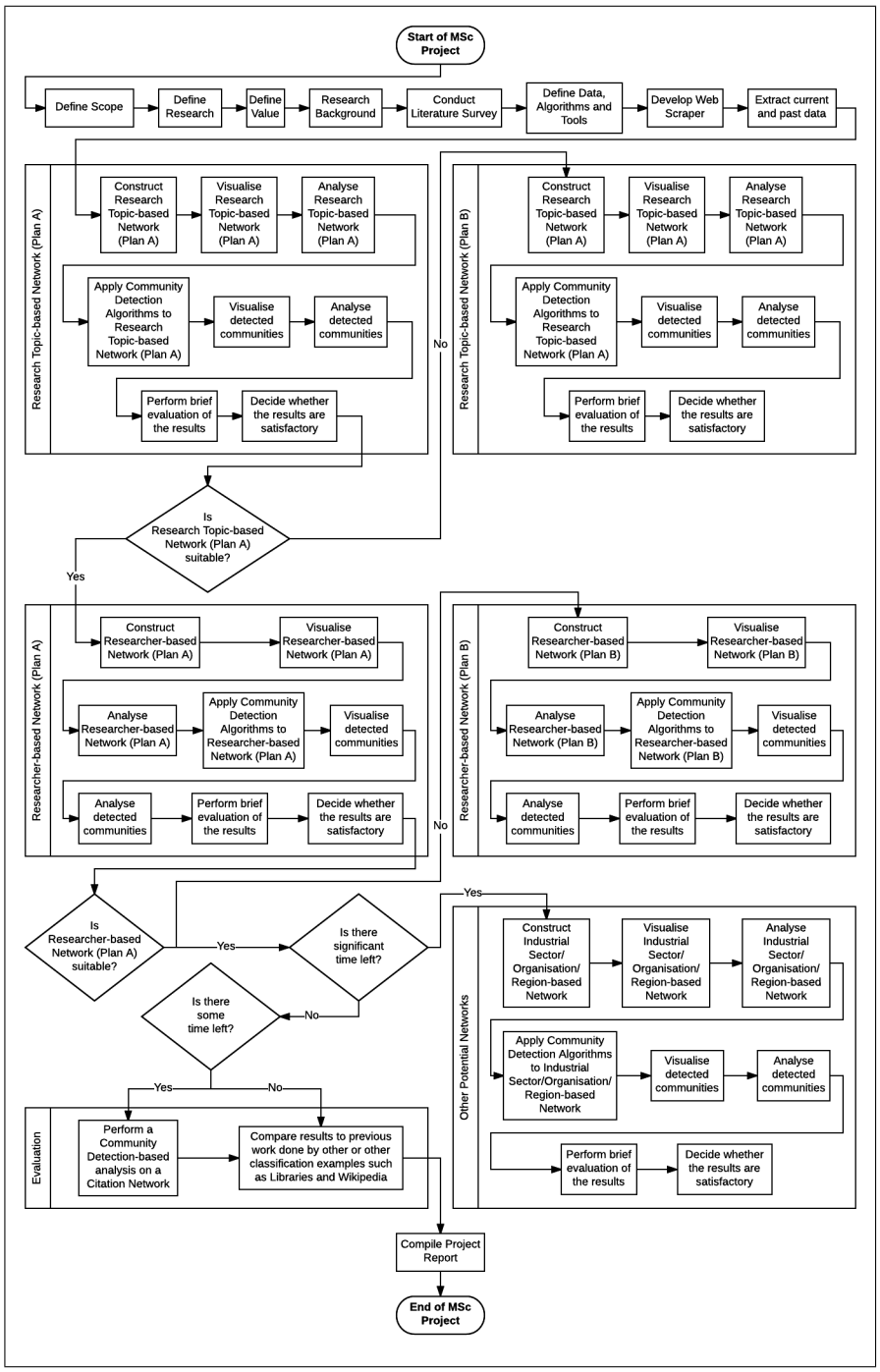


Figure 4.1: Flow chart created as part of the Project Plan, showing the transition between the different stages of the project.

5 Methodology

This chapter extends the Methodology section in Chapter 4: Project Design. Firstly, it provides an introduction to the EPSRC grant data used in this project and details its collection process. Secondly, the networks generated from the collected EPSRC data are introduced. Thirdly, the tools used in this project are presented and their use is highlighted. Furthermore, the contrast between grants as edges and grant records is outlined. Additionally, the process of formulating the node and edge attributes is presented, together with the normalisation process of the attribute values. Finally, the common experiment settings employed in all experiments are specified and the experiment process in which several edge weight interpretations and community detection algorithms are interchangeably tested is briefly described.

5.1 EPSRC grant data

This project uses data provided publicly by EPSRC through the *EPSRC GoW* service. It consists of current and historical data stored within the *Current* and *Past Grant Portfolio*, respectively.

5.1.1 EPSRC GoW service

The *GoW* service is a web-based facility providing information about research grants funded by EPSRC. The service is updated frequently, and consists of large amounts of information regarding historical and current grants, researchers, panels and quarterly summaries. It also includes search functionality allowing users to search the Web database.

5.1.2 EPSRC Current and Past Grant Portfolio

The *Current* and *Past Grant Portfolio* are sub-facilities of the *GoW* service providing access to current and historical grant data. Both facilities provide the same kind of information, however the access to information differs slightly. The *Past Grant Portfolio* requires a time period to be supplied, and provides grants based on it. This can be a start date, an end date or a start and end

date. Fig. 5.1 and Fig. 5.2 present the hierarchical structure of the EPSRC *Current* and *Past Grant Portfolio*, respectively.

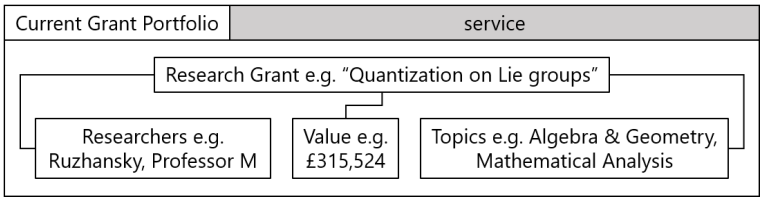


Figure 5.1: Hierarchical structure of the EPSRC Current Grant Portfolio.

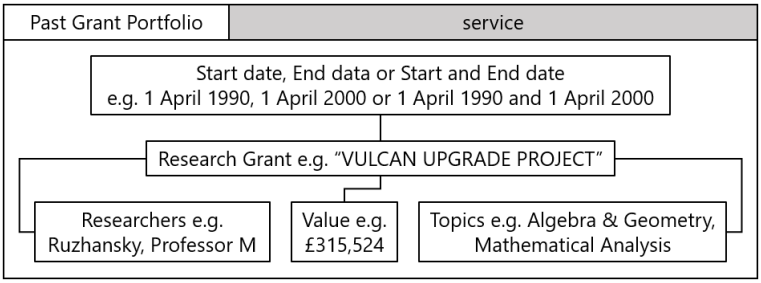
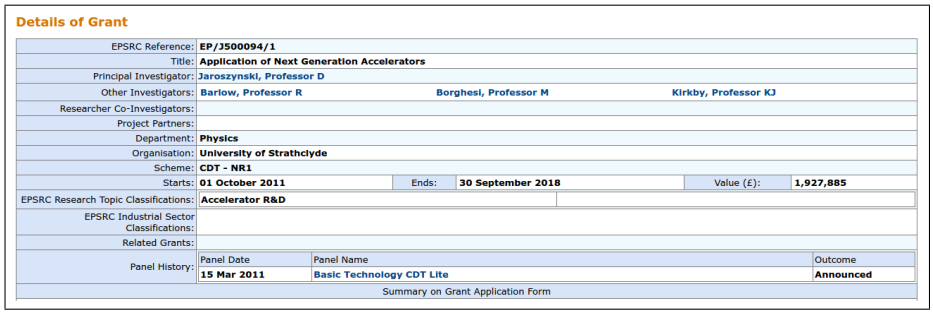


Figure 5.2: Hierarchical Structure of the EPSRC Past Grant Portfolio.

Each grant record is stored within a separate web page and contains details about the grant such as *reference*, *investigators (researchers)*, *partners*, *department*, *organisation*, *start* and *end date*, *value* and *topic* and *industrial sector classifications*. Fig. 5.3 shows an example of a grant record within the *EPSRC GoW* service.



Details of Grant

EPSRC Reference:	EP/J500094/1				
Title:	Application of Next Generation Accelerators				
Principal Investigator:	Jaroszynski, Professor D				
Other Investigators:	Barlow, Professor R	Borghesi, Professor M	Kirkby, Professor KJ		
Researcher Co-Investigators:					
Project Partners:					
Department:	Physics				
Organisation:	University of Strathclyde				
Scheme:	CDT - NR1				
Starts:	01 October 2011	Ends:	30 September 2018	Value (£):	1,927,885
EPSRC Research Topic Classifications:	Accelerator R&D				
EPSRC Industrial Sector Classifications:					
Related Grants:					

Panel History:	Panel Date	Panel Name	Outcome
	15 Mar 2011	Basic Technology CDT Lite	Announced

Summary on Grant Application Form

Figure 5.3: Grant record in the *EPSRC GoW* service.

The researchers within each grant record are linked to separate researcher records. Each researcher record is also stored within a separate web page and contains details about the researcher including *name*, *organi-*

sation, *department*, *current topics* and *grants*. Fig. 5.4 shows an example of a researcher record within the *EPSRC GoW* service.

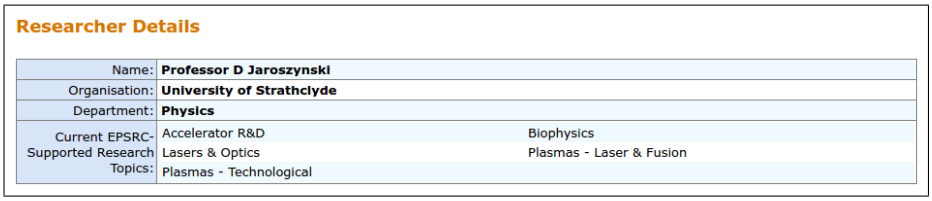


Figure 5.4: Researcher record in the *EPSRC Grants on Web (GoW)* service.

5.2 Collection of data from EPSRC

The previous sections provided an introduction to the data provided by EPSRC, the *Grants on the Web service (GoW)*, the *Current* and *Past Grant Portfolios* and the *Grant* and *Researcher* records. This project solely uses data collected from the *Current* and *Past Grant Portfolios*, and does not use any of the other information provided through the *GoW* service. In terms of the *Current* and *Past Grant Portfolios*, this project only makes use of data extracted from grant and researcher records.

Firstly, data within the following fields of a grant record was extracted: *EPSRC Reference*, *Principal Investigator*, *Other Investigators*, *Starts*, *Ends* and *EPSRC Research Topic Classifications*. Secondly, data within the *Name* and *Current EPSRC-Supported Research Topics* fields of a researcher record was extracted. Table 5.1 presents the number of current and historical EPSRC grant and researcher records from which data was collected.

Table 5.1: Number of current EPSRC grant and researcher records and historical grant records from which data was collected.

	Current	**Historical**	
		1990-2000	**2000-2010**
Number of Grant records	3175	17861	18692
Number of Researcher records	5874	10820	13385

Furthermore, the current and historical data used in this project is organised into two main data sets, while the historical data set is further divided into two sub-data sets as follows:

- **Current data set**, consisting of current (post 2010 start date) grant records collected on 8 July 2016
- **Historical data set**, consisting of grant records from two time periods:
  - **1990 to 2000**, consisting of grant records which started and ended between 1 April 1990 and 1 April 2000
  - **2000 to 2010**, consisting of grant records which started and ended between 1 April 2000 and 1 April 2010

As previously mentioned, both grant and researcher records are stored as separate web pages within the *GoW* service. In order to extract content from a web page, a technique called *web scraping* is employed. This is usually achieved by either using a third-party web scraper or by developing a web scraper from scratch, which extracts specified data from the underlying tags within the HTML code. In this case, a web scraper was developed using the *requests* and *lxml* Python libraries.

Essentially, the web scraper connects to the URL of each grant and researcher record and extracts the text found under specific fields such as *Reference*, *Principal Investigator* and *Other Investigators*, *Value* and *EPSRC Research Topic Classifications*. Once extracted, the data is validated and then comma-delimited and text qualified using quotation marks which preserves its original format and allows easy manipulation of it. Subsequently, it is stored within comma-separated files categorised by content and data set.

5.3 Generation of networks from EPSRC data

The data extracted from the *GoW* service was used to construct several networks of topics and researchers structured as follows:

1. **Networks of Topics, with nodes representing topics**
   - 1.1. **Topic-grant network**, with edges representing grants;
     - 1.1.1. current version using the current data set (2010-2016)
     - 1.1.2. historical version using historical data set (2000-2010)
     - 1.1.3. historical version using historical data set (1990-2000)
   - 1.2. **Topic-researcher network**, with edges representing researchers;
     - 1.2.1. current version using the current data set (2010-2016)[1]
2. **Networks of Researchers, with nodes representing researchers**
   - 2.1. **Researcher-grant network**, with edges representing grants;
     - 2.1.1. current version using the current data set (2010-2016)
     - 2.1.2. historical version using historical data set (2000-2010)
     - 2.1.3. historical version using historical data set (1990-2000)
   - 2.2. **Researcher-topic network**, with edges representing topics;
     - 2.2.1. current version using the current data set (2010-2016)[1]

5.4 Tools used in this project

During this study, several tools were employed in order to accomplish various development and analysis activities. This section describes the tools and comments on their use throughout the project.

5.4.1 JetBrains PyCharm for programming

JetBrains PyCharm [14] is an Integrated Development Environment (IDE) for programming in Python. It also provides support for writing code in Bash.

[1]The Topic (Nodes as Topics, Edges as Researchers) and Researcher network (Nodes as Researchers, Edges as Topics) were created using the Current data set only because Researcher records only consist of a researcher's current topics, and not their historical topics.

It was used to write the development code in primarily Python but also in Bash.

5.4.2 Microsoft Excel for data storage

Microsoft Excel [15] is a spreadsheet software featuring calculation, graphing tools, pivot tables, and macro programming support in Visual Basic. It was used to explore, filter and validate the collected data.

5.4.3 iGraph for network analysis and visualisation

iGraph [16] is a library collection for creating and manipulating graphs and analyzing networks. This tool was used to compute network properties, apply community detection algorithms, and produce visualisations of the networks created.

5.4.4 NetworkX for visualising adjacency matrices

NetworkX [17] is a package for the creation, manipulation, and study of the structure, dynamics and functions of complex networks in Python. It was used to visualise adjacency matrices.

5.4.5 Wordle for creating word cloud visualisations

Wordle [18] is a tool for visualising text as word clouds. By default, it computes each word's frequency and displays the more frequent words in a larger font than less frequent ones. Additionally, Wordle's advanced mode allows keeping words together, specifying a weight which controls the font size, as well as specifying a colour for each word. It was to create word cloud representations of topic clusters using several attributes to control font size and colour.

5.4.6 Adobe Photoshop for image editing

Adobe Photoshop [19] is a graphics editor developed by Adobe Systems. It was used for image editing as well as turning network plots produced using iGraph into complete network visualisations.

5.5 Common tasks

A number of tasks were carried out for every network constructed using every data set, therefore, the tasks are considered common and are described in this section. They include identifying the contrast between grants as edges and grant record, the formulation of node and edge attributes and the normalisation of attribute values.

5.5.1 Contrast between grants as edges and grant records

It is essential to point out that there is a clear difference between grants represented by edges and the actual grant records that appear on the *GoW* service. First and foremost, the former is not unique within a network, while the latter is unique within the *GoW* service. This is due to the way links between topics or researchers were established in a network. If a pair of topics co-appear in X grants, the two will in theory be linked by M edges. In this study, the M edges are merged into a single edge between the two topics, and the weight of the edge is assigned as the total sum of the X grants. Essentially, if a grant is associated with N topics, N(N-1)/2 edges are created to link the N topics into a fully connected mesh graph. For example, if grant Y has 4 topics, 6 edges are created linking the topics into a 4-clique graph. Furthermore, it is also crucial to specify that unlike grants, a network consists of nodes representing unique topics.

This causes issues when questions like the following are posed: *How many grants are in a specific community within a network's community structure? What is the value of the grants?* It is important to mention that it is not known which grant is represented by an edge. It is only known that it represents one or more grants depending on the edge weight. If this was known, providing an answer to the above questions becomes significantly easier. In contrast, the answer was achieved through a lengthier process. Furthermore, the edges linking topics within the same community and the edges linking topics from different communities were known. This greatly aided the identification process of both the number and value of grants within communi-

ties and between communities.

Firstly, the origin and destination topics of an edge within a network were retrieved. Secondly, during the data collection process, a data structure was created which stored the reference of each grant record and the topics that classify it. Next, a check was carried out against all grant records which verified whether both the retrieved origin and destination topics appeared under a grant record. Number and value counters were created to keep track of the cumulative number and value of grants. If the check was true, the reference and value of the grant were added as the key and value to a Python dictionary, in which all keys must be unique. Finally, at the end of the check, the number of grant references was counted and the grant values were summed up. This provided an answer to the questions asked, as the number and value of grants within a community was successfully identified. Additionally, using the same technique, the computation of the number and value of grants between communities and within the entire network was also achieved. The function written in order to achieve this task is presented in Code snippet C.1, part of Appendix C.

5.5.2 Formulation of node and edge attributes

Every network constructed using the data collected from EPSRC consists of at least one node and edge attribute, while others consist of two attributes. Both *Topic-grant* and *Researcher-grant* networks consist of a *number* and *value* attribute, while the *Topic-researcher* and *Researcher-topic* networks only consist of a *number* attribute. Visually, the network attributes control both the size of the node circle and the thickness of the edge line.

5.5.2.1 Number of grants/topics/researchers

The number attribute has a number of different contexts, depending on the network. Firstly, in the *Topic-grant* and *Researcher-grant* networks, the node number attribute represents the number of grants that a topic or researcher appears in. In the same networks, the edge number attribute represents the

number of grants two topics or researchers have in common, meaning they both appear within the same grant record. Secondly, in the *Topic-researcher* network, the node number attribute represents the number of researchers that a topic appears within, while the edge number attribute represents the number of researchers two topics have in common, meaning they both appear within the same researcher record. Thirdly, in the *Researcher-topic* network, the node number attribute represents the number of topics a researcher currently has, while the edge number attribute represents the number of topics two researchers have in common.

5.5.2.2 Value of grants

Similarly, the value attribute also has a number of different contexts, depending on the network. Firstly, networks that are not based on grant data do not contain of the value attribute. Secondly, in the *Topic-grant* and *Researcher-grant* networks, the node value attribute represents the value of a topic or researcher. This represents the value of the grants that contain that specific topic or researcher. In the same network, the edge value attribute represents the value of the grants that two topics or researcher have in common, meaning they both appear within the same grant record.

5.5.3 Normalisation of node and edge attribute values

The numerical values used as the node and edge attributes, especially the value attribute, represent significantly large numbers which cause issues in terms of development, analysis and visualisation. To accommodate this, the values underwent a normalisation process which scaled the value range down. The function written in order to achieve this task can be viewed in Code snippet C.2, part of Appendix C.

Furthermore, the formula used to normalize the values is presented below:

$$(val - old\_min) \times new\_range / old\_range) + new\_min) \tag{5.1}$$

where:

- *val* is the value being normalised
- *old_min* is the minimum of the initial value range
- *new_range* is the new range values will be normalised to
- *old_max* is the maximum of the initial value range
- *new_min* is the minimum of the new value range

5.6 Common experiments settings

Extensive comparison experiments were conducted with the purpose of identifying an optimal solution for the data used in this project. An optimal solution is represented by a combination of an edge weight interpretation and a community detection algorithm. All candidates were tested on all networks, interchangeably. This means that each community detection algorithm was applied to each network constructed using each edge weight interpretation. The testing criteria is initially based on modularity score and community size. Later, the actual clusters are manually evaluated based on coherence and balance. In total, three different edge weight interpretations and eight different community detection algorithms are considered.

5.6.1 Edge weight interpretations

Previously, the process carried out to normalise the values used as node and edge attributes was described. Depending on the network tested, one or both edge attributes formulated are used in the experiments. Three edge weight candidates are considered as follows:

1. edge weight, interpreted as **unweighted**
2. edge weight, interpreted as **the normalised number of grants**
3. edge weight, interpreted as **the normalised value of grants**

Note: the edge weight candidates may be occasionally abbreviated as **uw**, **wnn** and **wnv**, throughout this report.

Network that are not grant-based such as the *Topic-researcher* and *Researcher-topic* networks only consist of the number attribute, therefore, the experiments on those networks only involve the first two edge weight candidates. Furthermore, the edge weight attribute plays an important role in the clustering performance of the community detection algorithms, which is the reason why it is essential to conduct an experiment considering different edge weights in order to identify an optimal one for the data in question.

5.6.2 Community detection algorithms

In the Background chapter, the notion and multitude of community detection algorithms was introduced. In this project, eight different community detection algorithms are considered as candidates in the experiments:

1. **Infomap** by Rosvall and Bergstrom
2. **Spinglass** by Reichardt and Bornholdt
3. **Louvain** by Blondel et al.
4. **Label Propagation** by Raghavan et al.
5. **Leading Eigenvector** by Newman
6. **Walktrap** by Pons and Latapy
7. **Fast Greedy** by Newman et al.
8. **Edge Betweenness**

Each algorithm has already been tested by its authors and others on several real and artificial networks. In another project, this would make the experiments unnecessary. However, neither algorithm has been applied on the EPSRC data used in this project. Therefore, it is crucial to compare the performance of each algorithm in order to prove its suitability for the EPSRC current and historical data sets.

5.6.2.1 Spinglass by Reichardt and Bornholdt

Spinglass is another modularity optimisation proposed by *Reichardt and Bornholdt* [20] which is based on a combination between a popular statistical mechanic model called Potts spin glass, and the network community structure. The algorithm applies the technique of simulated annealing on Potts in order to achieve an optimal modularity [21].

5.6.2.2 Louvain by Blondel et al.

Louvain is a modularity optimisation algorithm introduced by *Blondel et al.* [22]. It proposes a two-phase hierarchical agglomerative approach which is an improvement of Fast Greedy by Newman et al. [23] The first phase of the algorithm involves the application of a greedy optimisation in order to detect communities. In the second phase, a new network is constructed using the communities found during the first phase as nodes. Edges between communities are represented as self-loops, while edges within communities are summed and represented as edges between the new nodes. This process is repeated until a single community remains [21].

5.6.2.3 Fast Greedy by Newman et al.

FastGreedy, an algorithm developed by *Newman et al.* [23] is based on a greedy optimisation method also applied to a hierarchical agglomerative approach. Initially, each node represents its own community. The communities are merged by the algorithm step by step until only one remains, containing all nodes. The greedy approach is applied at each step, by considering the largest increase or smallest decrease in modularity as the criteria for merging. Due to the hierarchical nature of the algorithm, it produces a hierarchy of community structures, and the comparison of modularity values determines the best community structure [21].

6 Networks of Topics

The idea behind a network of topics involves two ways that the network can be constructed. Likewise, the topics within a network can be analysed from the perspective of grants as well as researchers. Technically speaking, nodes in the network will represent topics regardless of perspective. However, edges can represent either grants or researchers. In the end, two different networks of topics are constructed, the *Topic-grant* network and the *Topic-researcher* network.

6.1 Topic-grant network

The *Topic-grant* network consists of nodes representing topics and edges representing grants. The *EPSRC Research Topic Classifications* field in each grant record consists of one or more topics that classify the grant. Only grant records with two or more topics were included in the analysis. Subsequently, a link between each topic and all other topics within each grant record was established. The link signifies the grant record that the topics all have in common, and is represented as an edge in a network. Fig. 6.1 provides a visual explanation of how the *Topic-grant* network was constructed using the collected EPSRC data, including the formulated node and edge attributes.

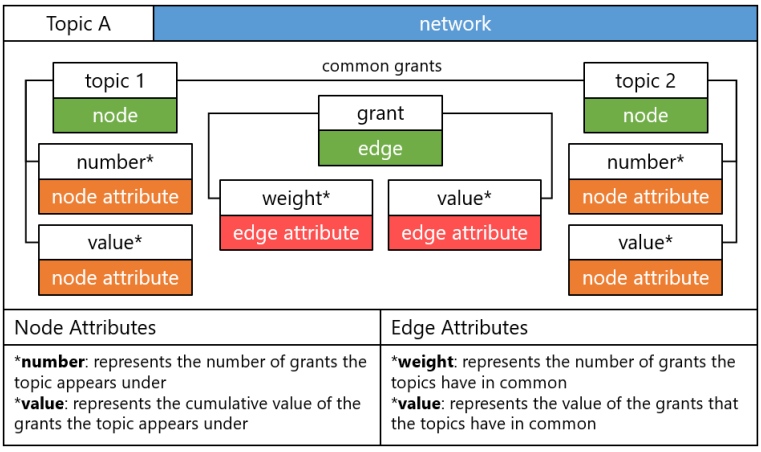


Figure 6.1: Visual explanation of how the *Topic-grant* network was structured and constructed using the collected EPSRC data, including the formulated node and edge attributes.

6.1.1 Node and edge attributes

The *Topic-grant* network contains two different node and edge attributes, the number and value of grants. Node and edge attributes are common between networks. Consequently, their formulation is considered a common task, and therefore it is described in Chapter 5: Methodology.

6.1.2 Properties of the Topic-grant network

The *Topic-grant* network constructed using the current (2010-2016) data set consists of 223 nodes representing as many topics and 2,008 edges representing common grants between topics. Table 6.1 presents the properties of both the historical and current *Topic-grant* networks.

Table 6.1: Properties of the *Topic-grant* network constructed using both the historical (1990 to 2010, 2000 to 2010) and current (2010 to 2016) data sets.

	1990-2000	**2000-2010**	**2010-2016**
Nodes	136	208	223
Edges	748	3592	2008
Type	Undirected	Undirected	Undirected
Weighted	Yes	Yes	Yes
Average Degree	11	34.538	18.009
Average Weighted Degree	12.721	35.337	19.543
Diameter	6	5	5
Density	0.081	0.167	0.081
Modularity	0.4	0.271	0.373
Average Clustering Coefficient	0.453	0.59	0.597
Average Path Length	2.54	2.077	2.395

In comparison, both historical networks contain less nodes, with the *2000 to 2010* and *1990 to 2010* networks consisting of 208 and 136 nodes, respectively. This is expected, as the number of research topics in the past was lower, and gradually increased over time. Interestingly, the *2000 to 2010* network consists of more edges which means more topics are connected through common grants. However, the increased number of edges also seems to correlate with the fact that the network is unconnected, while the other networks are connected. A network is unconnected when an edge

does not exist between every pair of nodes. Furthermore, all three networks are weighted and in the creation of Table 6.1 and Fig. 6.2, the number of grants edge weight attribute is used.

6.1.3 Visualisation of Topic-grant network

A visualisation of the *Topic-grant* network, presented in Fig. 6.2, was produced using iGraph. It features nodes in blue, and edges in grey. The size of the node circle represents the number of grants node attribute. The width of the edge line represents the number of grants edge attribute. The topic(s) that appear in the highest number of grant records are coloured in red.

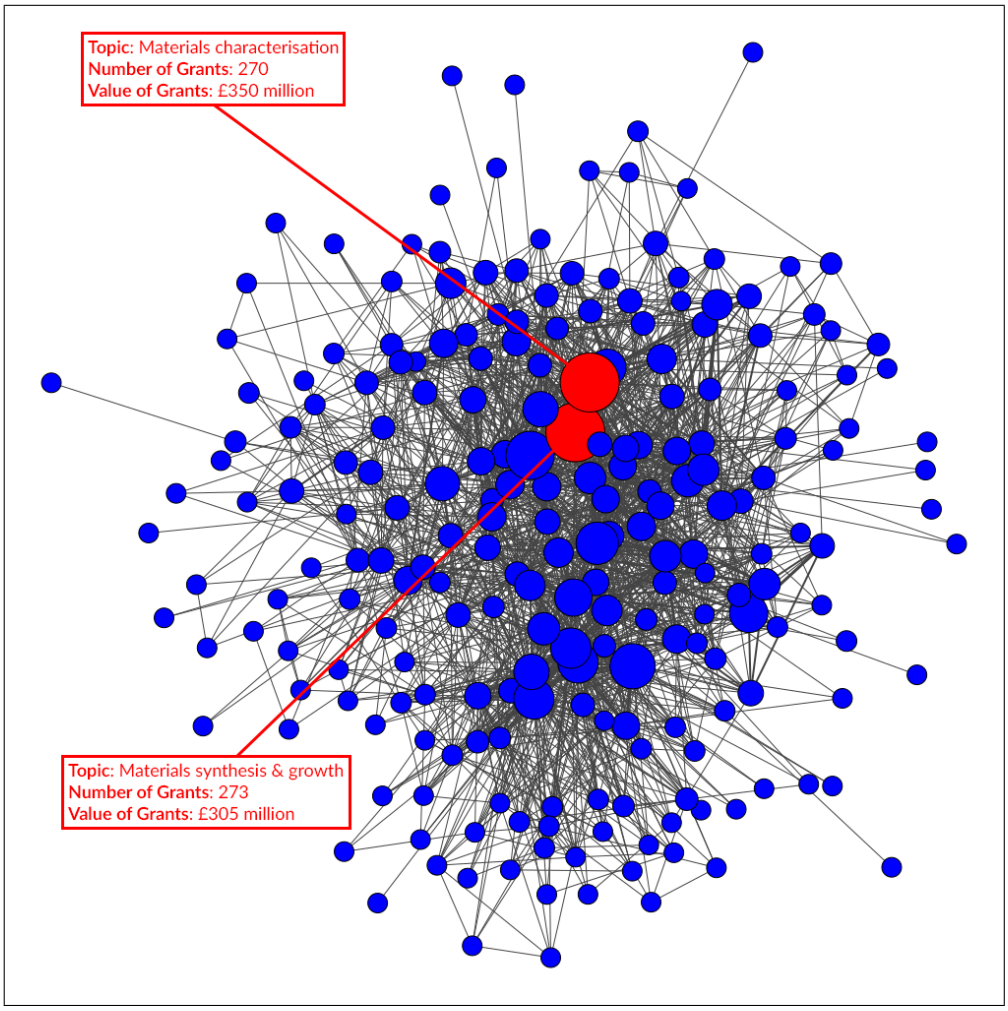


Figure 6.2: Visualisation of *Topic-grant* network constructed using the current (2010 to 2016) data set. The topic(s) that appear in the highest number of grant records are coloured in red.

6.2 Topic-researcher network

The *Topic-researcher* network consists of nodes representing topics and edges representing researchers. The *Current EPSRC-Supported Research Topics* field in each researcher record consists of one or more topics that classify the grants supported by EPSRC that the researcher is currently an investigator in. Only researcher records with two or more topics were included in the analysis. Subsequently, a link between each topic and all other topics within each researcher record was established. The link signifies the researcher record that the topics all have in common, and is represented as an edge in a network. Fig. 6.3 provides a visual explanation of how the *Topic-researcher* network was constructed using the collected EPSRC data, including the formulated node and edge attributes.

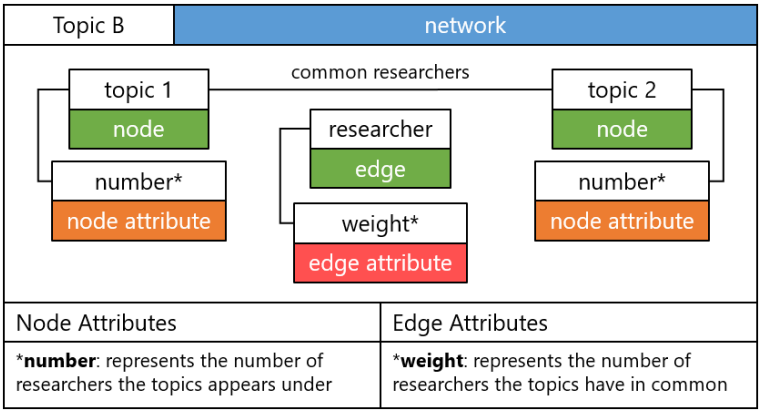


Figure 6.3: Visual explanation of how the *Topic-researcher* network was structured and constructed using the collected EPSRC data, including the formulated node and edge attributes.

6.2.1 Node and edge attributes

The *Topic-researcher* network contains one node and edge attribute, the number of grants. In the *Topic-researcher network* edges represent researchers, therefore, it does not contain the value of grants attribute. Node and edge attributes are common between networks. Consequently, their formulation is considered a common task, and therefore it is described in Chapter 5: Methodology.

6.2.2 Properties of Topic-researcher network

The *Topic-researcher* network represents the second interpretation of the topic data, and consists of topics represented by 225 nodes connected by 5,192 edges. Table 6.2 presents the properties of the *Topic-researcher* network.

Table 6.2: Properties of the *Topic-researcher* network constructed using the current (2010 to 2016) data set.

	2010-2016
Nodes	225
Edges	5192
Type	Undirected
Weighted	Yes
Average Degree	46.151
Average Weighted Degree	52.436
Diameter	4.0
Density	0.206
Modularity	0.234
Average Clustering Coefficient	0.715
Average Path Length	1.925

There is a substantial increase in the number of edges compared to the *Topic-grant* network. This increase is partly due to the fact that the number of researcher records (5,874) considered in the analysis exceeds the number of grant records (3,175) considered by 2,699 records. Furthermore, this network is also weighted, but this time, the edge weight represents the number of researchers two topics have in common. Moreover, the *Topic-researcher network* was only constructed using the current (2010-2016) data set because researcher records only consist of the current topics of a researcher. This limits the comparison of the *Topic-researcher* and *Topic-grant* networks, as they can only be contrasted based on the current (2010-2016) data set.

Furthermore, it is essential to indicate the slight difference in the number of nodes between the *Topic-grant* network (223 nodes) and *Topic-researcher* network (225 nodes). The former considers grants with two or more topics while the latter considers researchers with two or more topics. In either network, records may exist where a specific topic appears in a single record and

is also the single topic within that record. This means that the topic will not be considered in the analysis, as links to other topics cannot be established.

6.2.3 Visualisation of Topic-researcher network

A visualisation of the *Topic-researcher* network, presented in Fig. 6.4, was produced using iGraph. It features nodes in blue, and edges in grey. The size of the node circle represents the number of researchers node attribute. The width of the edge line represents the number of researchers edge attribute. The topic(s) that appear in the highest number of researcher records are coloured in red.

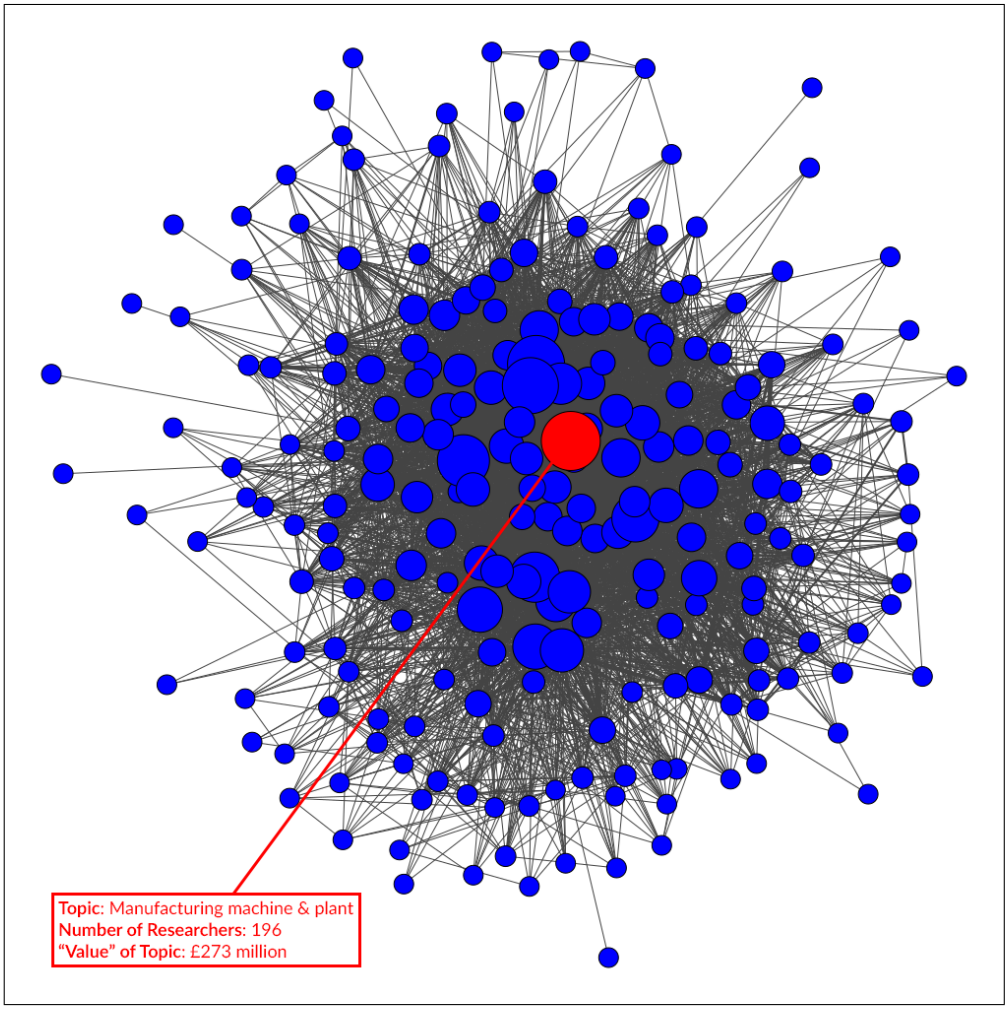


Figure 6.4: Visualisation of *Topic-researcher* network constructed using the current (2010 to 2016) data set. The topic(s) that appear in the highest number of researcher records are coloured in red.

7 Clustering of Topics

This chapter describes the process of clustering the *Topic-grant* and *Topic-researcher* networks. The experiments carried out are detailed, while the results are presented. Additionally, results are evaluated and discussed through comparisons to historical data and between the two Topic networks.

7.1 Clustering of Topic-grant network

The process of clustering the *Topic-grant* network involves a number of different stages including carrying out experiments, producing results, evaluating results as well as conducting comparative analysis.

7.1.1 Experiment

An experiment was carried out in order to identify an optimal edge weight and community detection algorithm. The full settings of the experiment are outlined and detailed in Chapter 5: Methodology.

The experiment was divided into a number of phases. In the first phase, a number of community detection algorithms were applied to the current *Topic-grant* network constructed using each of the edge weight. The resulting modularity scores of the identified community structure and the number of generated communities were compared across all edge weight and community detection algorithm candidates. Most community detection algorithms make use of the edge weight attribute in the clustering process, which meant that the *unweighted* edge weight had a low performance and was excluded from the experiment in the first phase. Furthermore, a number of community detection algorithms such as *Label Propagation* and *Edge Betweenness* were also excluded due to low performance. In contrast, the *Louvain*, *Spinglass* and *Fast Greedy* algorithms remained to be tested further in the second phase. Table 7.1 presents the results produced during the first phase of the experiment. The edge weights and community detection algorithms considered for the second phase of the experiment are in **red**.

Table 7.1: Number of communities identified (left value) and the modularity score of the community structure discovered (right value) as a result of applying several community detection algorithms to the *Topic-grant* network constructed using the current (2010 to 2016) data set. Three of the column names are abbreviated: **uw**, **wnn** and **wnv** stand for unweighted, weighted by normalized number of grants and weighted by normalized value of grants, respectively. High-performance candidates are in **red**.

	uw		**wnn**		**wnv**	
Infomap	3	0.004	9	0.332	11	0.377
Spinglass	6	0.355	5	0.375	5	0.392
Louvain	5	0.347	6	0.373	6	0.385
Label Propagation	1	0.0	1	0.0	1	0.0
Leading Eigenvector	5	0.312	5	0.302	10	0.311
Walktrap	5	0.279	19	0.295	24	0.328
Fast Greedy	5	0.314	4	0.359	5	0.369
Edge Betweenness	164	0.038	166	0.042	105	0.15

During the second phase, the number of topics clustered within each community was compared across all experiment candidates, and it was decided that all remaining edge weights and community detection algorithms should proceed to the final testing phase, in order to be tested further. Table 7.2 presents the results produced in the second phase of the experiment.

Table 7.2: Number of topics clustered within each community discovered as a result of applying several community detection algorithms to the Topic-grant network constructed using the current data set (2010 to 2016). Six of the column names are abbreviated: **C1** stands for Community 1, **C2** stands for Community 2 and so on. Two of the row names are abbreviated: **wnn** and **wnv** stand for weighted by normalized number of grants and weighted by normalized value of grants, respectively.

		C1	**C2**	**C3**	**C4**	**C5**	**C6**
wnn	**Spinglass**	61	35	37	30	60	-
	Louvain	29	61	63	10	34	26
	Fast Greedy	35	84	66	38	-	-
wnv	**Spinglass**	63	17	51	63	29	-
	Louvain	46	9	29	61	43	35
	Fast Greedy	24	75	69	33	22	-

In the third phase and the final phase, each community of topics identified using each one of the edge weights and community detection algorithms were manually compared to each other, side by side, with coherence as the decision criteria. Table 7.3 presents an example of how the final phase of the experiment was conducted on the *Topic-grant* network.

Table 7.3: Example of how the experiment on the *Topic-grant* network constructed using the current (2010-2016) data set was conducted. In the example, the topics in **red** within the **Spinglass - wnn** column are not in the **Louvain - wnn** and vice versa.

Community 1 (Spinglass - wnn)	**Community 1 (Louvain - wnn)**
ageing: chemistry/biochemistry	ageing: chemistry/biochemistry
algebra & geometry	analytical science
analytical science	bioelectronic devices
bioelectronic devices	bioinformatics
bioinformatics biological	medicinal chem.
biological & medicinal chem.	**biomaterials**
biomechanics & rehabilitation	biomechanics & rehabilitation
biomedical neuroscience	biomedical neuroscience
...	...

The experiment concluded that the most rational clustering of topics was produced using the edge weight, *weighted by the normalized number of grants* and the *Louvain* community detection algorithm. Consequently, an optimal combination of edge weight and community detection algorithm was identified.

7.1.2 Results

The application of the optimal solution identified on the *Topic-grant* network resulted in the identification of 6 different communities of topics. Table 7.4 presents the number of nodes representing topics, the number and value of grants and the predominant words within each community discovered in the current *Topic-grant* network. The complete clustering of the historical and current *Topic-grant* networks is presented in Tables B.1, B.2 and B.3, part of Appendix B.

Table 7.4: Number of nodes and grants, value of grants and the predominant words based on word frequency of each community identified in the *Topic-grant* network constructed using the current (2010 to 2016) data set. The number of grants includes duplicate grants, as a grant can be part of more than one community. Subsequently, the value of grants also includes the duplicate grants. However, the last column represents the number and value of unique grants in communities within the current *Topic-grant* network.

	Number (topics)	Number (grants)	Value (grants)	Predominant words based on frequency of words
C1	29	511	£629M	biology, biomedical, science
C2	61	774	£862M	design, computing, psychology
C3	63	1338	£1.5B	chemistry, engineering, materials
C4	10	317	£332M	mathematical, analysis
C5	34	480	£584M	management, engineering, energy
C6	26	484	£766M	optical, devices, quantum
All	223	3072	£3.5B	engineering, biology, chemistry

Community 1 (Biology, Science) has a clear focus surrounding biology, chemistry, medicine and science and does not consist of any topics that would be irrational to be clustered as part of it. Topics clustered within this community include: *ageing: chemistry/biochemistry*, *biomedical sciences* and *drug formulation & delivery*. By far, the topic receiving most funding is *med.instrument.device& equip.*, valued at £192M. This value is justified as the topic also appears in 155 grants, more than any other topic in Community 1. Figure 7.1 presents a word cloud representation of Community 1.

Figure 7.1: Topics clustered within Community 1 in the *Topic-grant* network. Font size represents the number of grants that each topic appears in.

Community 2 (IT, Psychology, Criminology) is not as well defined as Community 1 because it consists of several topics which have obvious differences such as *product design*, *artificial intelligence*, *developmental psychology*, *human communication in ict*, *criminal law & criminology* and *comput./corpus linguistics*. This contrast is justified considering the significant size of the community. Furthermore, there are grants classified by two or more topics which are different, in theory, such as *artificial intelligence* and *linguistics*, for example. However, *Natural language processing* is a field of both *artificial intelligence* and *computational linguistics*. Figure 7.2 presents a word cloud representation of Community 2.

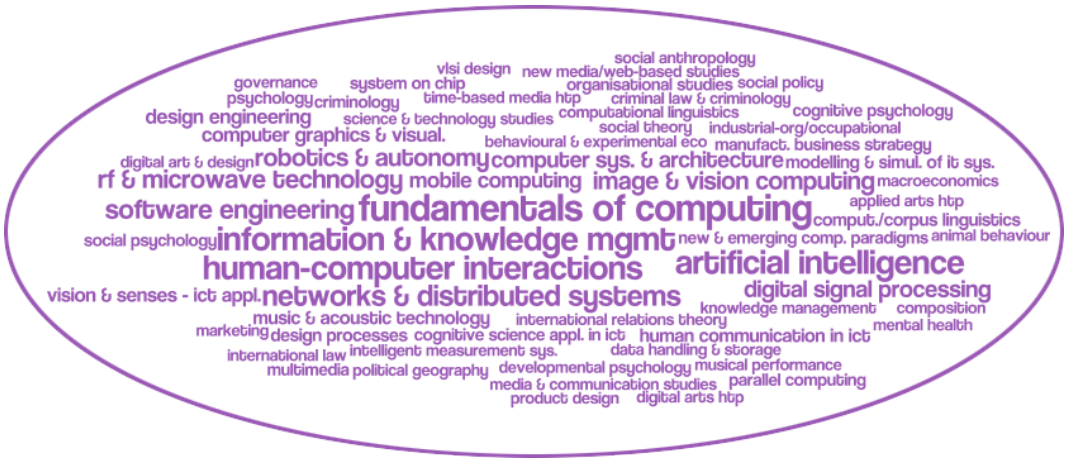


Figure 7.2: Topics clustered within Community 2 in the *Topic-grant* network. Font size represents the number of grants that each topic appears in.

Community 3 (Energy, Power) also represents a comprehensible clustering enclosing topics such as *fluid dynamics*, *aerodynamics*, *bioenergy*, *microsystems*, *wind power* and *energy - marine & hydropower*. It consists of three major topics both in terms of number of grants that each topic appears in but also the value of those grants: *materials characterisation* (270 grants, worth £350M), *materials synthesis & growth* (273 grants, worth £305M) and *manufacturing machine & plant* (196 grants worth £273M). Figure 7.3 presents a word cloud representation of Community 3.

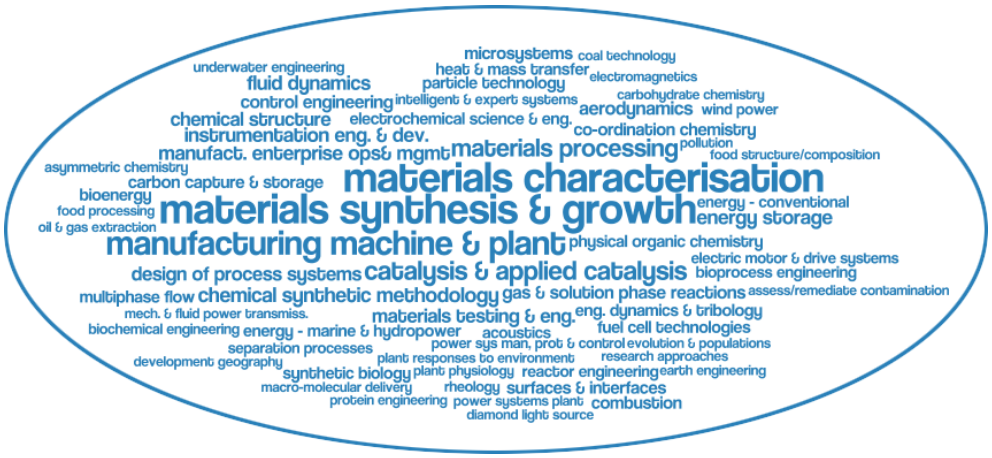


Figure 7.3: Topics clustered within Community 3 in the *Topic-grant* network. Font size represents the number of grants that each topic appears in.

Community 4 (Mathematics) is the smallest community in size (only 10 topics) but also the most well-defined community, as the topics clustered within it have a concrete, shared focus in Mathematics. It consists of research topics including *algebra & geometry*, *mathematical physics* and *continuum mechanics*. *Algebra & geometry* and *statistics & appl. probability* are the topics which appear in the highest number of grants, 131 and 120, respectively. In terms of value, the latter is valued higher than the former, £200M compared to £69M. Figure 7.4 presents a word cloud representation of Community 4.

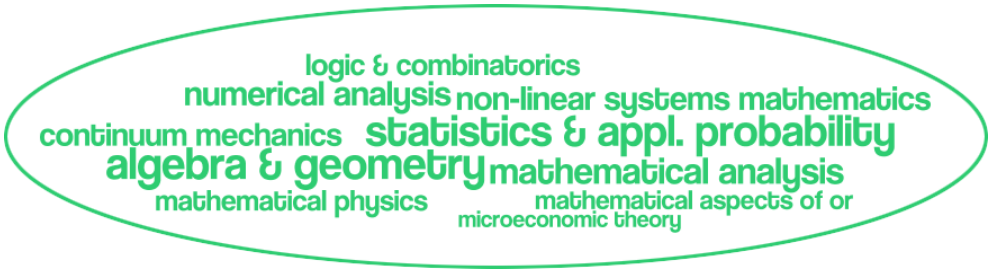


Figure 7.4: Topics clustered within Community 4 in the *Topic-grant* network. Font size represents the number of grants that each topic appears in.

Community 5 (Environment) represents a coherent clustering of topics including *energy efficiency*, *geohazards*, *environment & health* and *urban & land management*. This community represents topics from different fields which appear in grants that share an aim in tackling an environment-related problem such as climate change or global warming. The most popular topic

in terms of number of grants is *energy efficiency*, with 100 grants valued at £150M. Figure 7.5 presents a word cloud representation of Community 5.

Figure 7.5: Topics clustered within Community 5 in the *Topic-grant* network. Font size represents the number of grants that each topic appears in.

Community 6 (Physics, Electricity) is the last identified community and another community which consists of a rational clustering of topics surrounding the fields of *physics* and *electricity*. Topics include *solar technology*, *biophysics* and *electronic devices & subsys.*. *Condensed Matter Physics* is the topic that appears in in the highest number of grants, 82, worth £97M. Figure 7.6 presents a word cloud representation of Community 6.

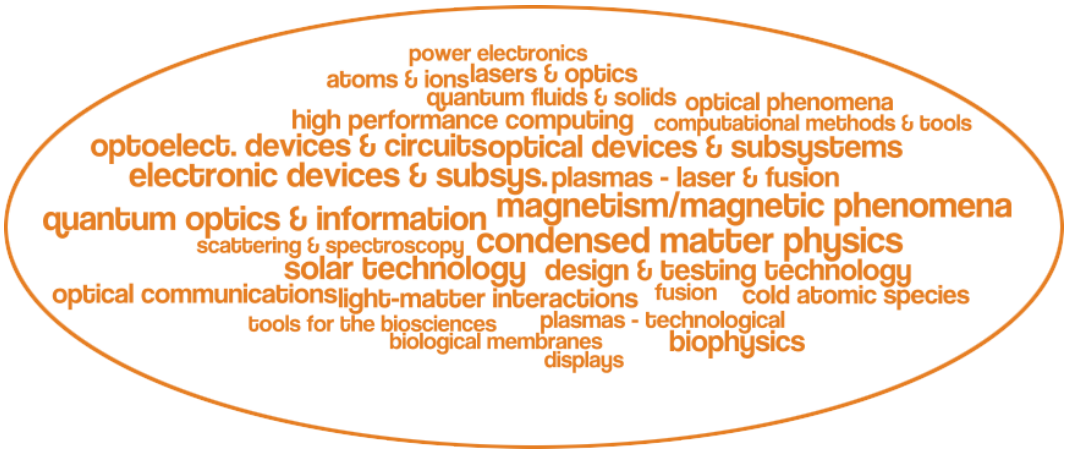


Figure 7.6: Topics clustered within Community 6 in the *Topic-grant* network. Font size represents the number of grants that each topic appears in.

Click **here** to view a full, high-resolution word cloud representation.

7.1.2.1 Visualisation of community structure

A visualisation of the community structure discovered in the current *Topic-grant* network by the optimal solution identified, presented in Fig. 7.7, was produced using iGraph. It features nodes in 6 different colours depending on which community they are clustered in. Edges between the nodes in each community are grey, while edges between nodes from different communities are excluded altogether. The size of the node circle represents the number of grants node attribute. The width of the edge line represents the number of grants edge attribute. The topics with the highest number of grants are identified by a darker shade of the colour assigned to their community.

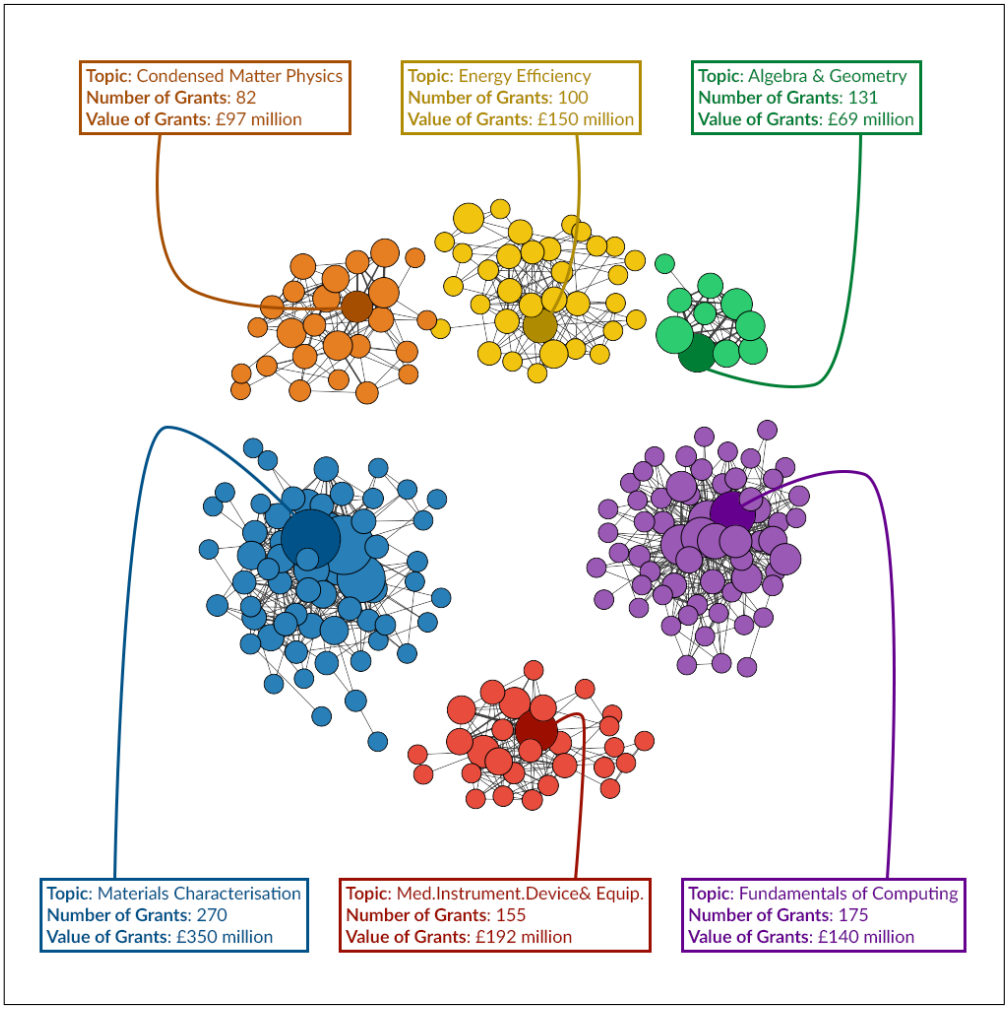


Figure 7.7: Visualisation of the community structure within the *Topic-grant* network constructed using the current (2010 to 2016) data set. The topics with the highest number of grants are identified by a darker shade of the colour.

7.1.3 Evaluation

A strong network clustering means that a node pair within the same cluster will have a higher similarity compared to a node pair consisting of nodes from two different clusters. The purpose of the evaluation phase is to determine whether in reality this is also true.

First and foremost, the origin and destination nodes of each edge in the *Topic-grant* network were identified. This forms a pair of nodes as follows: (origin node, destination node). Certain edges link nodes that are in the same cluster, while others link nodes that are in different clusters. Therefore, some node pairs represent nodes from the same cluster, while others represent nodes from different clusters. Subsequently, the Dice and Jaccard similarity between each node pair was calculated. Finally, in order to obtain an overall perspective of the similarity of nodes within and between clusters, the average Dice and Jaccard similarity was calculated.

Indeed, the results show that for both the current and historical Topic-grant networks, nodes within the same cluster have a higher similarity than nodes from different clusters. Table 7.5 presents the results of the evaluation phase carried out on the Topic-grant network constructed using both the historical (1990 to 2010) and current (2010 to 2016) data sets.

Table 7.5: Dice and Jaccard similarity coefficients of node pairs within and between clusters in the Topic-grant network constructed using both the historical (1990 to 2010) and current (2010 to 2016) data sets. Each node pair represents an edge which links two nodes from the same cluster or two different clusters. **IN** stands for within communities, while **OUT** means between communities.

	1990-2000	**2000-2010**	**2010-2016**
Node pairs IN	437	1940	1122
Node pairs OUT	311	1652	886
Average Dice similarity IN	0.465	0.510	0.428
Average Dice similarity OUT	0.346	0.433	0.354
Difference between IN and OUT	0.119	0.077	0.074
Average Jaccard similarity IN	0.316	0.356	0.286
Average Jaccard similarity OUT	0.217	0.283	0.220
Difference between IN and OUT	0.099	0.073	0.066

7.1.4 Discussion

The *Topic-grant* network was constructed using both the historical (1990 to 2010) and current (2010 to 2016) data sets. This means a comparative analysis can be conducted comparing the data sets in terms of research trend and funding, and clustering.

7.1.4.1 Comparison to historical data

Over the years, the research trend led to new topics being defined, while others were discontinued. For example, *bionanoscience*, *escience* and *language acquisition* are topics which existed *from 2000 to 2010* but do not exist currently. In contrast, *animal organisms*, *political geography* and *ageing: chemistry/biochemistry* are some of the current topics that were defined after 2010.

Similarly, the funding trend also evolved as more recent grants saw a significant increase in the funding support provided. Currently, there are 3,072 grants within communities with a total value of £3.5B. Between 2000 and 2010, researchers worked on 16,617 grants, valued at £4.9B. These figures indicate a significant difference in the number of grants. However, this is justified, as the two time periods compared are not equal, as the former covers 6 years of grants, while the latter covers 10 years. More importantly, the difference in value is not considerable, which shows the progress of research funding over the years, as current grants receive significantly more funding than they would have 10 years ago.

Furthermore, this is also supported by the number and value of grants completed between 1990 and 2000. Researchers worked on a slightly less number of grants than between 2000 and 2010, but they also received significantly less funding, £1.7B. Tables 7.6 and 7.7 present the number of nodes representing topics, the number and value of grants and the predominant words within each community in the *Topic-grant* network constructed using the historical (2000 to 2010) data set and historical (1990 to 2000) data set, respectively.

Table 7.6: Number of nodes and grants, value of grants and the predominant words based on word frequency of each community identified in the *Topic-grant* network constructed using the historical (2000 to 2010) data set. The number of grants includes duplicate grants, as a grant can be part of more than one community. Subsequently, the value of grants also includes the duplicate grants. However, the last column represents the number and value of unique grants in communities within the historical *Topic-grant* network.

	Number (topics)	**Number (grants)**	**Value (grants)**	**Predominant words based on word frequency**
C1	67	8682	£2.6B	chemistry, biology, science
C2	43	5167	£1.3B	engineering, mathematical
C3	25	1394	£699M	energy, power
C4	2	1	£1M	science
C5	71	4099	£1.3B	design, arts, digital
All	208	16617	£4.9B	engineering, biology, design

In terms of clustering, the historical communities identified within the *Topic-grant* network hold slight differences when compared to the current communities. First and foremost, the community detection algorithm identified 5 communities in both historical (1990 to 2000, 2000 to 2016) networks. This decrease may signify the result of the contrast in the number of topics and the actual topics between the current and historical networks. Moreover, the most well-defined, Community 4 (Mathematics) identified in the current *Topic-grant* network, is not as well-defined anymore in the historical network, as it is part of a larger community in Community 2 (Engineering, Mathematics). This symbolises the current increase in the number of grants that are focused on *mathematics* only, rather than a combined effort including other topics such as *engineering*. Furthermore, between 2000 and 2010, only one grant was classified simultaneously by *soil science* and *crop science*. In the current (2010 to 2016) data set, *crop science* is not present anymore. Perhaps, its removal could be justified by the similarity between the two topics, deeming one of them as unnecessary. This also indicates a potential reason why the two topics did not form a community in the current *Topic-grant* network.

Table 7.7: Number of nodes and grants, value of grants and the predominant words based on word frequency of each community identified in the *Topic-grant* network constructed using the historical (1990 to 2000) data set. The number of grants includes duplicate grants, as a grant can be part of more than one community. Subsequently, the value of grants also includes the duplicate grants. However, the last column represents the number and value of unique grants in communities within the historical *Topic-grant network*.

	Number (topics)	**Number (grants)**	**Value (grants)**	**Predominant words based on word frequency**
C1	28	2015	£246M	engineering, energy, management
C2	25	3995	£661M	optical, devices, materials
C3	17	1328	£88M	mathematical, analysis
C4	39	3455	£496M	engineering, ict, design
C5	27	2567	£385M	chemistry, catalysis, energy
All	136	12791	£1.7B	engineering, chemistry, systems

7.1.4.2 Motivation for the Topic-researcher network

The *Topic-researcher* network represents an alternative and another way of interpreting the topic data to construct a network. An alternative represents a different perspective on the data. The *Topic-grant* network is one way of analysing topics, from the point of view of grants. The *Topic-researcher* network is a second way of analysing topics, from the point of view of researchers. Grant records and researcher records are significantly different, and one may provide different or better results than the other. Furthermore, having a different way of analysing the data means that a comparative analysis of the results can be carried out. Therefore, both networks are considered important due to the different and valuable insights that can be translated from the data.

7.2 Clustering of Topic-researcher network

The process of clustering the *Topic-researcher* network involves a number of different stages including carrying out experiments, producing results, evaluating results as well as conducting comparative analysis.

7.2.1 Experiment

In order to ensure a consistent comparative analysis between the two Topic networks, the optimal solution identified as a result of the experiment on the *Topic-grant* network, is also considered the optimal solution for the *Topic-researcher* network. However, the experiment was still carried out and the results are presented in Tables 1.31-1.33, part of the Supplementary material.

7.2.2 Results

The application of the optimal solution identified on the *Topic-researcher* network resulted in the identification of 4 different communities of topics, 2 less than within the *Topic-grant* network. Note that the *Topic-researcher* network was constructed using the current (2010 to 2016) data set only, as researcher records within EPSRC only provide the current topics of a researcher. Table 7.8 presents the number of nodes representing topics and the predominant words within each community discovered in the *Topic-researcher* network. The complete clustering of the *Topic-researcher* network is presented in Table B.4, part of Appendix B.

Table 7.8: Number of nodes and the predominant words based on word frequency of each community identified within the *Topic-researcher* network constructed using the current (2010 to 2016) data set.

	Number (topics)	**Predominant words based on frequency of words**
C1	39	engineering, management
C2	62	psychology, design
C3	15	mathematical, analysis
C4	109	engineering, chemistry
All	225	engineering, management, science

7.2.2.1 Visualisation of community structure

Similarly to the *Topic-grant* network, a visualisation of the community structure discovered in the current *Topic-researcher* network was produced and is presented in Fig. 7.8.

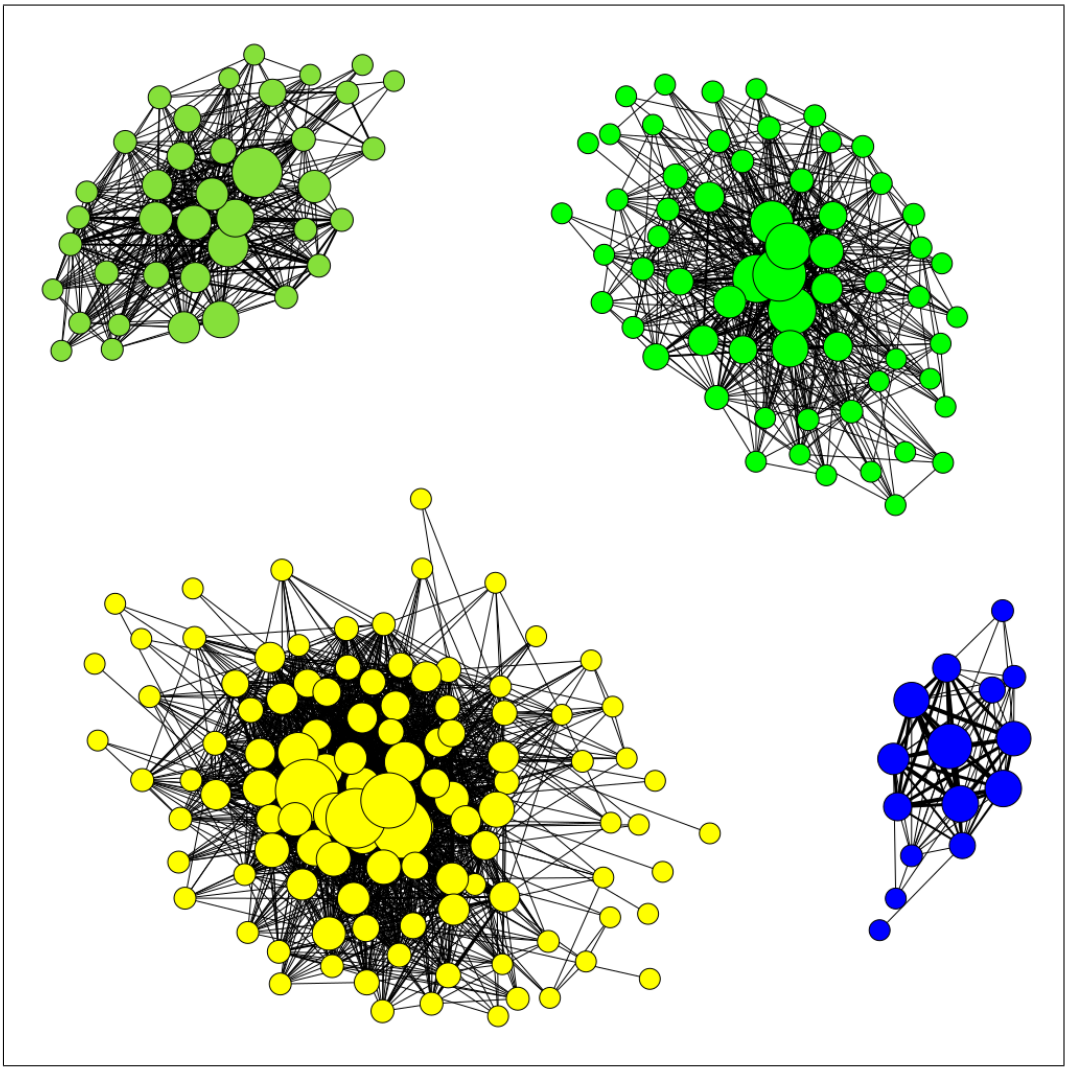


Figure 7.8: Visualisation of the community structure within the Topic-researcher network constructed using the current (2010 to 2016) data set.

7.2.3 Evaluation

Similarly to the *Topic-grant* network, the *Topic-researcher* network also underwent the evaluation phase and the results show that for both the current and historical *Topic-researcher* networks, nodes within the same cluster have a higher similarity than nodes from different clusters. Table 7.9 presents the

results of the evaluation phase carried out on the *Topic-researcher* network constructed using the current (2010 to 2016) data set.

Table 7.9: Dice and Jaccard similarity coefficients of node pairs within and between clusters in the *Topic-researcher* network constructed using the current (2010 to 2016) data set. Each node pair represents an edge which links two nodes from the same cluster or two different clusters. **IN** stands for within communities, while **OUT** means between communities.

	2010-2016
Node pairs IN	2921
Node pairs OUT	2271
Average Dice similarity IN	0.543
Average Dice similarity OUT	0.517
Difference between IN and OUT	0.026
Average Jaccard similarity IN	0.393
Average Jaccard similarity OUT	0.359
Difference between IN and OUT	0.034

7.2.4 Discussion

The *Topic-grant* and *Topic-researcher* networks represent two interpretations of the topic data. This means a comparative analysis can be conducted comparing the two networks in terms of clustering.

7.2.4.1 Comparison to Topic-grant network

There are obvious differences between the clustering produced using the *Topic-grant* network and *Topic-researcher* network. Firstly, the number of communities identified differs, as using the former 6 communities were identified, compared to 4 when using the latter. This results in an imbalance in community size, with one of the communities (Community 4) identified in the *Topic-researcher* network consisting of 109 topics. A large community is also a broad community, which means that is less specific and lacks the capability to represent one or more clear research areas. It also causes other communities to be significantly smaller in size.

Furthermore, the *Mathematics* community (Community 4) identified using the *Topic-grant* network is larger in size and its previously well-defined structure is ”harmed” by the irrational addition of topics such as *genomics* when identified using the *Topic-researcher* network. Moreover, the clustering produced using the *Topic-researcher* network also failed to identify the *Biology* community (Community 1) which was successfully identified using the *Topic-grant* network. That being said, there are also similarities between the communities identified using the two networks, as the *Engineering* and *Chemistry* communities appear in the community structure of both networks.

In conclusion, it is concluded that the clustering produced using the *Topic-grant* network is more coherent and balanced than the one produced using the *Topic-researcher* network.

8 Networks of Researchers

The idea behind a network of researchers involves two ways that the network can be constructed. Likewise, the researchers within the network can be analysed from the perspective of grants as well as topics. Technically speaking, nodes in the network will represent researchers regardless of perspective. However, edges can represent either grants or topics. In the end, two different networks of researchers are constructed, the *Researcher-grant* network and the *Researcher-topic* network.

8.1 Researcher-grant network

The *Researcher-grant* network consists of nodes representing researchers and edges representing grants. The *Principal* and *Other Investigators* fields in each grant record consist of one or more researchers that collaborate on the grant. Only grant records with two or more researchers were included in the analysis. Subsequently, a link between each researcher and all other researchers within each grant record was established. The link signifies the grant record that the researchers all have in common, and is represented as an edge in a network. Due to the significantly large size of the *Researcher-grant* network, the network had to be sampled with the sample consisting of nodes connected by edges with an edge weight of 2 or more. Fig. 8.1 provides a visual explanation of how the *Researcher-grant* network was constructed using the collected EPSRC data, including the formulated node and edge attributes.

8.1.1 Node and edge attributes

The *Researcher-grant* network contains two different node and edge attributes, the number and value of grants. Node and edge attributes are common between networks. Consequently, their formulation is considered a common task, and therefore it is described in Chapter 5: Methodology.

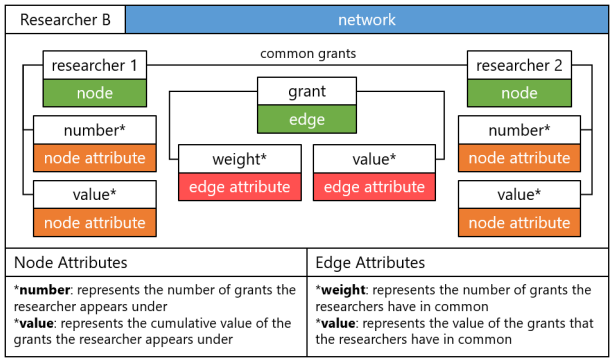


Figure 8.1: Visual explanation of how the *Researcher-grant* network was structured and constructed using the collected EPSRC data, including the formulated node and edge attributes.

8.1.2 Properties of Researcher-grant network

The *Researcher-grant* network is a carbon copy of the *Topic-grant* network, with the exception that, in this case, nodes represent researchers and not topics. Table 8.1 presents the properties of both the historical and current *Researcher-grant* networks.

Table 8.1: Properties of the *Researcher-grant* network constructed using both the historical (1990 to 2010) and current (2010 to 2016) data sets.

	1990-2000	**2000-2010**	**2010-2016**
Nodes	1847	2434	260
Edges	2002	4919	208
Type	Undirected	Undirected	Undirected
Weighted	Yes	Yes	Yes
Average Degree	2.168	4.042	1.60
Average Weighted Degree	2.798	7.794	2.592
Diameter	18.0	36.0	16.0
Density	0.001	0.002	0.006
Modularity	0.978	0.977	0.955
Average Clustering Coefficient	0.648	0.748	0.578
Average Path Length	3.676	6.874	1.942

The current *Researcher-grant* network is composed of 260 nodes and 208 edges representing one or more common grants between two researchers. The network features a significantly disconnected structure, shown in Fig. 8.2, which is primarily due to the rarity in repeated collaboration among researchers on EPSRC-supported grants, which may last up to 8 years.

Furthermore, both historical *Researcher-grant* networks consist of both an increased number of nodes and edges when compared to their current data equivalent. Furthermore, all three networks are weighted and in the in the creation of Table 8.1 and Fig. 8.2, the number of grants edge weight attribute is used.

8.1.3 Visualisation of Researcher-grant network

A visualisation of the *Researcher-grant* network, presented in Fig. 8.2, was produced. It features nodes in blue, and edges in grey. The size of the node circle represents the number of grants node attribute. The width of the edge line represents the number of grants edge attribute. The researcher(s) that appear in the highest number of grant records are coloured in red.

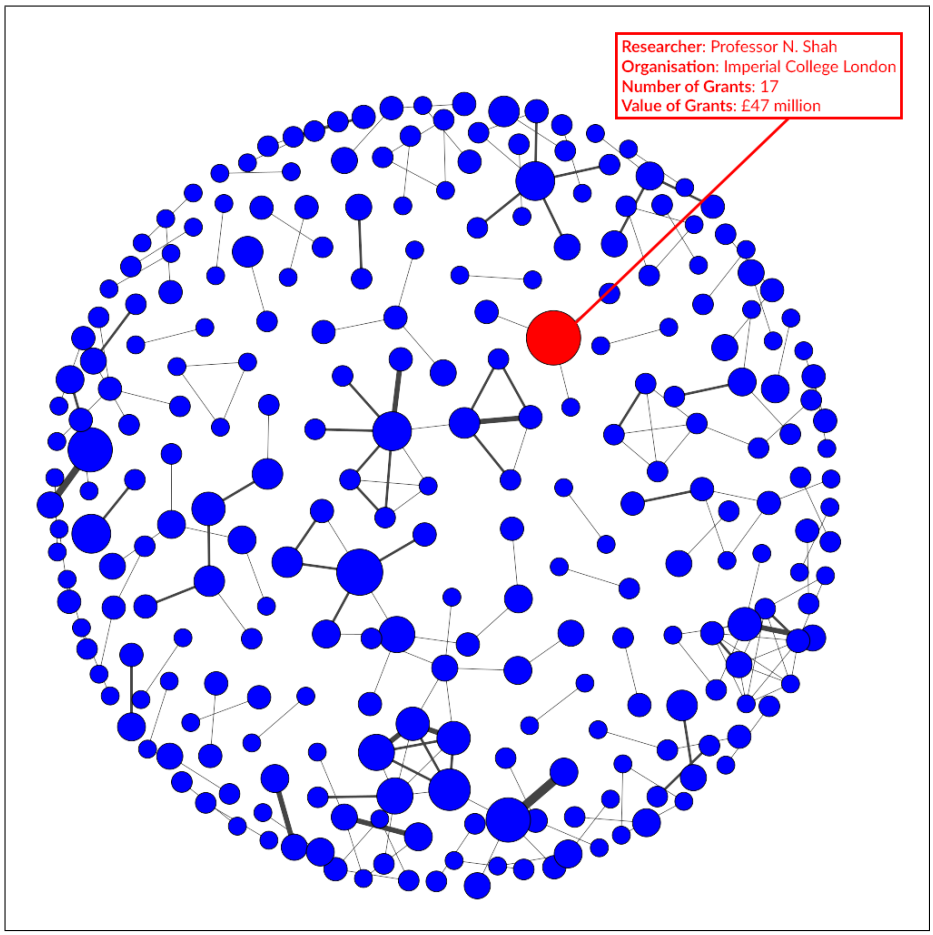


Figure 8.2: Visualisation of the *Researcher-grant* network constructed using the current (2010 to 2016) data set. The researcher(s) that appear in the highest number of grant records are coloured in red.

8.2 Researcher-topic network

The *Researcher-topic* network consists of nodes representing researchers and edges representing topics. The *Current EPSRC-Supported Research Topics* field in each researcher record consists of one or more topics that classify the grants supported by EPSRC that the researcher is currently an investigator in. Only researcher records with two or more topics were included in the analysis. Subsequently, each researcher record was compared to all other researcher records, and if they had at least one topic in common a link between the two researchers was established. The link signifies the topic(s) that two researchers have in common, and is represented as an edge in a network. Due to its significantly large size, the *Researcher-topic* network was sampled and only includes nodes connected by edges with an edge weight of 5 or more. Fig. 8.3 provides a visual explanation of how the *Researcher-topic* network was constructed using the collected EPSRC data, including the formulated node and edge attributes.

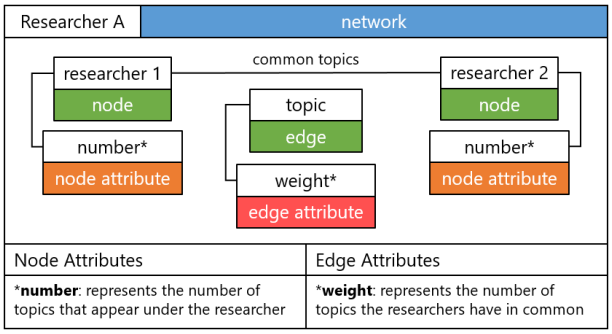


Figure 8.3: Visual explanation of how the *Researcher-topic* network was structured and constructed using the collected EPSRC data, including the formulated node and edge attributes.

8.2.1 Node and edge attributes

The *Researcher-topic* network contains one node and edge attribute, the number of topics. In the *Researcher-topic network* edges represent topics, therefore, it does not contain the value of grants attribute. Node and edge attributes are common between networks. Consequently, their formulation is considered a common task, and therefore it is described in Chapter 5: Methodology.

8.2.2 Properties of Researcher-topic network

The *Researcher-topic* network is a reversed version of the *Topic-researcher* network. It consists of 655 researchers linked by 4548 edges representing common topics between the researchers. Table 8.2 presents the properties of the *Researcher-topic* network.

Table 8.2: Properties of the *Researcher-topic* network constructed using the current (2010 to 2016) data set.

	2010-2016
Nodes	655
Edges	4548
Type	Undirected
Weighted	Yes
Average Degree	13.887
Average Weighted Degree	27.258
Diameter	15.0
Density	0.021
Modularity	0.738
Average Clustering Coefficient	0.825
Average Path Length	4.278

In comparison to the current *Researcher-grant* network which consists of 260 nodes and 208 edges, this network consists of substantially more nodes and edges. Due to its significantly large size, the *Researcher-topic* network had to be sampled with the sample consisting of nodes connected by edges with an edge weight of 5 or more. Furthermore, the network is undirected and weighted and in the creation of Table 8.2 and Fig. 8.4, the number of topics edge weight attribute is used.

8.2.3 Visualisation of the Research-topic network

A visualisation of the *Researcher-topic* network, presented in Fig. 8.4, was produced using iGraph. It features nodes in blue, and edges in grey. The size of the node circle represents the number of topics node attribute. The width of the edge line represents the number of topics edge attribute. The researcher(s) with the highest number of topics are coloured in red.

As illustrated in Fig. 8.4, the structure of the *Researcher-topic* network stands out from the others. Both *Topic* networks feature a conglomeration of nodes densely connected. In contrast, the *Researcher-grant* network consists the opposite of that, as it is sparsely connected with up to a maximum of 4 connected nodes. The *Researcher-topic* network represents a combination between the two. Based on an initial observation, an unsurprising guess would indicate that a community detection algorithm has already been applied to this network. However, it has not, as its structure clearly resembles the layout of collaboration between researchers. Several small groups of researchers, densely connected internally, but sparsely connected externally, can be spotted. This can also translate to the assumption that the communities represent common topic interests between researchers.

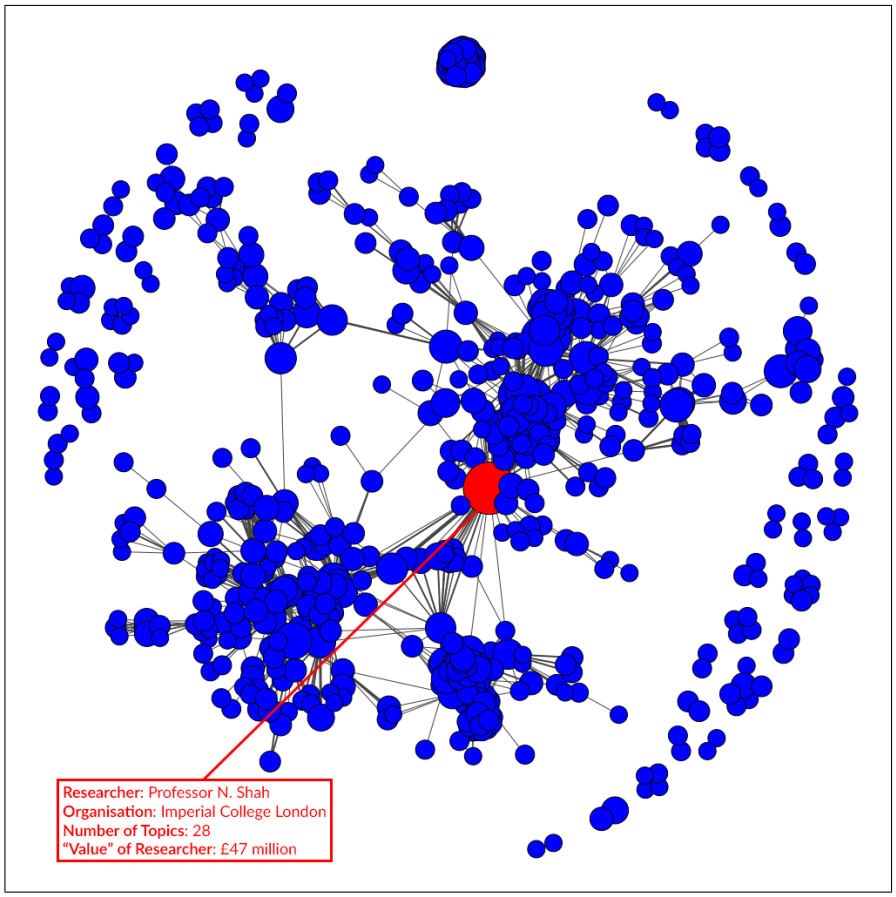


Figure 8.4: Visualisation of the *Researcher-topic* network constructed using the current (2010 to 2016) data set. The researcher(s) with the highest number of topics are coloured in red.

9 Clustering of Researchers

This chapter describes the process of clustering the *Researcher-grant* and *Researcher-topic* networks. The experiments are detailed, while the results are presented. Additionally, results are evaluated and discussed through comparisons to historical data and between the two *Researcher* networks.

9.1 Clustering of Researcher-grant network

The process of clustering the *Topic-grant* network involves a number of different stages including carrying out experiments, producing results, evaluating results as well as conducting comparative analysis.

9.1.1 Experiment

Due to the limited amount of time available and in order to ensure a consistent comparative analysis, the optimal solution identified as a result of the experiment on the *Topic-grant* network, is also considered the optimal solution for the *Researcher-grant* network. However, the experiment was still carried out and the results are presented in Tables 1.36-1.38, 1.43-1.45 and 1.49-1.51, part of the Supplementary material.

9.1.2 Results

The application of the optimal solution identified on the *Researcher-grant* network resulted in the initial and excessive identification of 89 communities of researchers, which signifies the lack of strong and frequent collaboration relationships between researchers. Table 9.1 presents the number of nodes representing researchers and the number and value of grants within each community discovered in the current *Researcher-grant* network.

Consequently, it seems that two or more researchers collaborating multiple times as part of a grant is a rarity within the collected EPSRC data. Due to the sparse nature of the communities identified, only the two largest communities are presented in the report.

Table 9.1: Number of nodes and grants and value of grants within the 2 largest communities identified in the *Researcher-grant* network constructed using the current data set (2010 to 2016). The number of grants includes duplicate grants, as a grant can be part of more than one community. Subsequently, the value of grants also includes the duplicate grants. However, the last column represents the number and value of unique grants in communities within the current *Researcher-grant* network.

	C1	**C2**	**Total**
Number of researchers	10	12	22
Number of grants	33	35	65
Value of grants	£103M	£87M	£177M

9.1.2.1 Visualisation of community structure

Similarly to the *Topic-grant* network, a visualisation of the community structure discovered in the current *Researcher-grant* network was produced and is presented in Fig. 9.1.

9.1.3 Evaluation

Similarly to the *Topic-grant* network, the *Researcher-grant* network also underwent the evaluation phase and the results show that for both the current and historical *Researcher-grant* networks, nodes within the same cluster have a higher similarity than nodes from different clusters. Table 9.2 presents the results of the evaluation phase carried out on the *Researcher-grant* network constructed using both the historical (1990 to 2000, 2000 to 2010) and current (2010 to 2016) data sets.

9.1.4 Discussion

Similarly to the *Topic-grant*, the *Researcher-grant* network was constructed using both the historical (1990 to 2010) and current (2010 to 2016) data sets. This means a comparative analysis can be conducted comparing the data sets in terms of research and funding trend, and clustering.

9.1.4.1 Comparison to historical data

Similarly to the historical data comparison of the *Topic* networks, there is a clear funding trend which increases more rapidly the more recent the grant

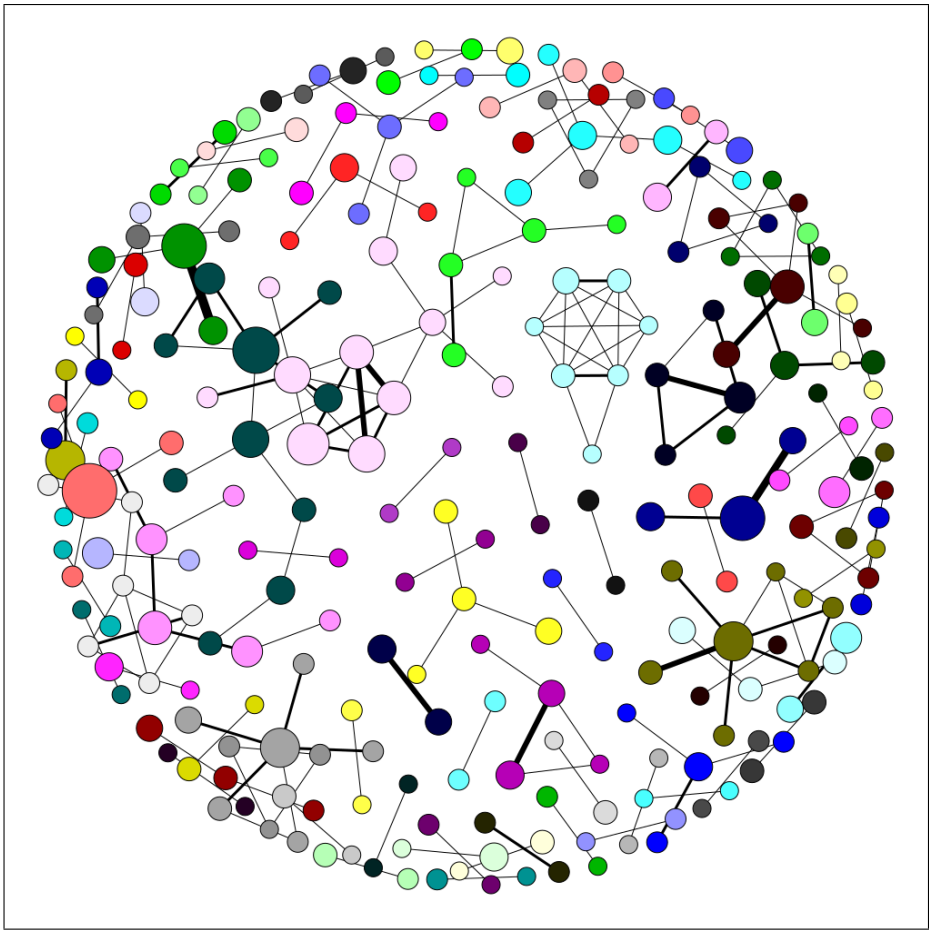


Figure 9.1: Visualisation of the community structure within the *Researcher-grant* network constructed using the current (2010 to 2016) data set.

Table 9.2: Dice and Jaccard similarity coefficients of node pairs within and between communities in the *Researcher-grant* network constructed using both the historical (1990 to 2010) and current (2010 to 2016) data sets. Each node pair represents an edge which connects two nodes within the same community or in two different communities. **IN** stands for within communities, while **OUT** means between communities.

	1990-2000	**2000-2010**	**2010-2016**
Node pairs IN	1992	4901	206
Node pairs OUT	10	18	2
Average Dice similarity IN	0.840	0.825	0.823
Average Dice similarity OUT	0.342	0.321	0.336
Difference between IN and OUT	0.481	0.504	0.505
Average Jaccard similarity IN	0.736	0.740	0.762
Average Jaccard similarity OUT	0.210	0.200	0.202
Difference between IN and OUT	0.526	0.540	0.560

is. Currently, within the two largest communities, there are 65 grants worth a total value of £177M. Table 9.3 presents the number of nodes representing researchers and the number and value of grants within each community in the *Researcher-grant* network constructed using the historical (2000 to 2010) data set.

Table 9.3: Number of nodes and grants and value of grants within the 4 largest communities identified in the *Researcher-grant* network constructed using the historical (2000 to 2010) data set. The number of grants includes duplicate grants, as a grant can be part of more than one community. Subsequently, the value of grants also includes the duplicate grants. However, the last column represents the number and value of unique grants in communities within the historical *Researcher-grant* network.

	C1	**C2**	**C3**	**C4**	**Total**
Number of researchers	49	46	55	65	215
Number of grants	213	184	208	278	866
Value of grants	£87M	£123M	£136M	£234M	£551M

From 2000 to 2010, within the four largest communities, researchers worked on 866 grants, valued at £551M. These figures indicate a significant difference (798) to the number of grants completed after 2010, within the two largest communities. However, this is justified, when considering the fact that the four largest communities in the historical (2000 to 2010) data set consist of 866 grants compared to 68 grants within the two largest communities in the current (2010 to 2016) data set. More importantly, the difference in value is not significant, as the grants in the current (2010 to 2016) data set are valued higher.

Only comparing the two largest communities in both the historical (2000 to 2010) and current (2010 to 2016) data sets, the number of grants in the former is significantly larger than the one in the latter, 213 grants compared to 33 and 184 grants compared to 35. Likewise, the value of the grants is also larger, but not significantly larger, £210M compared to £190M. This mirrors the insights discovered during the analysis of the *Topic-grant* network, as research funding increased over the years, with current grants receiving

significantly more funding than grants in the past.

Furthermore, this is supported by the number and value of the grants within the five largest communities in the historical (1990 to 2000) data set. Researchers worked on a slightly less number of grants than between 2000 and 2010, but they also received significantly less funding, £130M in total. Table 9.4 presents the number of nodes representing researchers and the number and value of grants within each community in the *Researcher-grant* network constructed using the historical data set (1990 to 2000).

Table 9.4: Number of nodes and grants and value of grants within the 5 largest communities identified in the *Researcher-grant* network constructed using the historical (1990 to 2000) data set. The number of grants includes duplicate grants, as a grant can be part of more than one community. Subsequently, the value of grants also includes the duplicate grants. However, the last column represents the number and value of unique grants in communities within the historical *Researcher-grant* network.

	C1	**C2**	**C3**	**C4**	**C5**	**Total**
Number of researchers	31	21	46	35	30	163
Number of grants	115	78	150	147	129	610
Value of grants	£34M	£18M	£21M	£41M	£17M	£130M

9.1.4.2 Motivation the Researcher-topic network

Similarly to the *Topic-researcher* network, the *Researcher-topic* network represents an alternative to the *Researcher-grant* network. Therefore, both Researcher networks are considered important due to the different and valuable insights that can be translated from the data.

9.2 Clustering of Researcher-topic network

The process of clustering the *Researcher-topic* network involves a number of different stages including carrying out experiments, producing results, evaluating results as well as conducting comparative analysis.

9.2.1 Experiment

Due to the limited amount of time available and in order to ensure a consistent comparative analysis, the optimal solution identified as a result of the experiment on the *Topic-grant* network, is also considered the optimal solution for the *Researcher-topic* network. However, the experiment was still carried out and the results are presented in Tables 1.55-1.57, part of Supplementary material.

9.2.2 Results

In contrast to the *Researcher-grant* network, the application of the optimal solution identified on the *Researcher-topic* network resulted in the less excessive initial identification of 49 communities of researchers. However, the number of identified communities is still too large and represents communities consisting of small numbers of researchers that have a small number of topics in common.

Furthermore, it seems that a large number of researchers having multiple common topics is a rarity within the EPSRC data. Due to the sparse nature of the communities identified, the eight largest communities are presented here. Table 9.5 presents the number of nodes representing researchers within the eight largest communities in the *Researcher-topic* network.

Table 9.5: Number of nodes within the eight largest communities identified in the *Researcher-topic* constructed using the current (2010 to 2016) data set.

	C1	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**	**Total**
Number of researchers	60	120	30	24	43	48	126	46	225

9.2.2.1 Visualisation of community structure

Similarly to the *Topic-grant* network, a visualisation of the community structure discovered in the current *Researcher-topic* network was produced and is presented in Fig. 9.1.

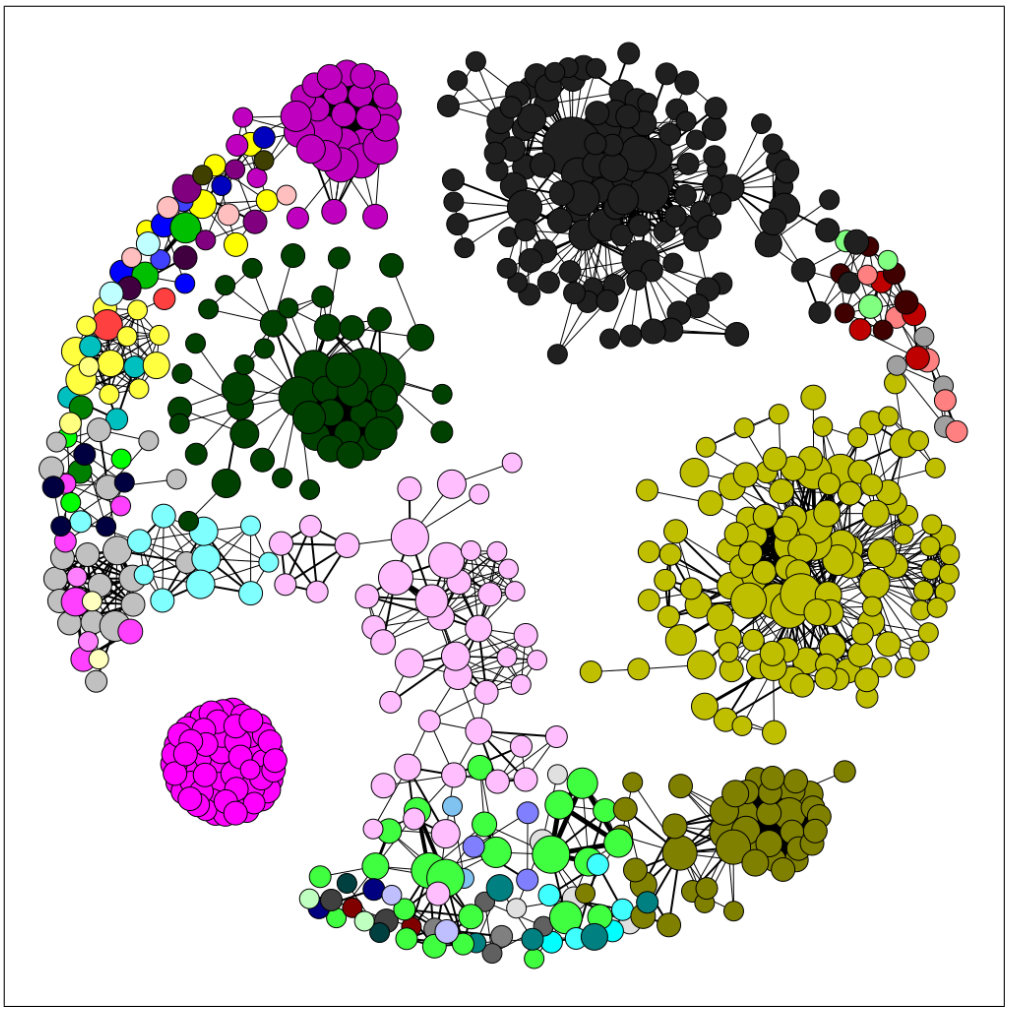


Figure 9.2: Visualisation of the community structure within the *Researcher-topic* network constructed using the current (2010 to 2016) data set.

9.2.3 Evaluation

Similarly to the *Topic-grant* network, the *Researcher-topic* network also underwent the evaluation phase and the results show that nodes within the same cluster have a higher similarity than nodes from different clusters. Table 9.6 presents the results of the evaluation phase carried out on the *Researcher-topic* network constructed using the current (2010 to 2016) data set.

Table 9.6: Dice and Jaccard similarity coefficients of node pairs within and between communities in the *Researcher-topic* network constructed using the current (2010 to 2016) data set. Each node pair represents an edge which connects two nodes within the same community or in two different communities. *IN* stands for within communities, while *OUT* means between communities.

	2010-2016
Node pairs IN	4273
Node pairs OUT	275
Average Dice similarity IN	0.835
Average Dice similarity OUT	0.405
Difference between IN and OUT	0.430
Average Jaccard similarity IN	0.779
Average Jaccard similarity OUT	0.273
Difference between IN and OUT	0.506

9.2.4 Discussion

The *Researcher-grant* and *Researcher-topic* networks represent two interpretations of the researcher data. This means a comparative analysis can be conducted comparing the two networks in terms of clustering.

9.2.4.1 Comparison to Researcher-topic network

The clustering results produced using both networks clearly indicate that the *Researcher-topic* network is the more rational and valuable interpretation of the researcher data, as expected prior to the creation process of the *Researcher* networks.

In one hand, there is a network that consists of researchers and links between them symbolising a work collaboration between the two. As proved by the results, this is fairly rare in the academic world. Unless two researchers have a relationship, are part of the same institution or organisation or are located in the same city, it is hard to believe that they would collaborate on multiple occasions. Moreover, it is essential to note that there is a clear difference between collaborating on a government-funded grant and any other research project. Certain grants can last up to 7 years and at the end of the grant the researchers that worked on it will most probably not collaborative

again for a lengthy period of time, if at all. In conclusion, the grant-based interpretation of the researcher data did not prove to be significantly valuable. However it did provide interesting insights into the progress of research and research funding over a 26-year long period.

On the other hand, there is a network that consists of researchers and links between them symbolising a shared research interest between the two. This type of connection between two researchers is not limited to a relationship, institution or organisation, or location. Only a shared research interest is required. In the academic world, this is very common as researchers from all over the world can connect and relate to each other through a common research field. The ideal result from such a network is a fairly low number of communities consisting of researchers that share an interest in a number of different topics. However, due to the large number of researchers, the number of identified communities turns out to be large and at the same time, sparse. That being said, this interpretation of the data is still a better option than the first, as it produces a more compelling clustering based on the researcher data available.

10 Discussion

In this chapter, the key results produced are summarised, and the limitations and improvements of the project are outlined, while the improvements are translated into future work, which is recommended.

10.1 Summary of key results

At the end of this thesis project, three key results were produced. Firstly, a rational, well-defined and balanced clustering of current and historical research topics was achieved. Secondly, a concrete work collaboration network between researchers was discovered, however, its clustering resulted in a substantial amount of communities which slightly diminished the extent of its analysis and its potential value.

Finally, an optimal solution composed of an edge weight and a community detection algorithm was identified. The optimal solution identified outperformed all other solutions considered in the comparison experiments. Furthermore, the key results produced represent the fulfilment of the objectives set at the start of the project.

Additionally, the impact of the results depends on who makes use of which results and how those results are used. Specifically, the identification of rational topic and researcher clusters is valuable to EPSRC as it represents a coherent way of clustering topics based on similarity and relatedness, a solution to a problem which currently exists.

Furthermore, the data analysis performed and the network visualisations produced represent invaluable insights into both the current and historical data. On the other hand, the identification of an optimal combination of edge weight and community detection algorithm as a result of the exhaustive experiments conducted, is more important to someone who wants to address a different but similar clustering problem as the one addressed in this project.

10.2 Limitations

Due to the short amount of time available, the analysis performed on the *Researcher* networks is slightly limited when compared to the analysis of the *Topic* networks, in terms of the quantity of data collected and the extent of the analysis. However, this was justified in the end, as the value and importance of the results produced during the analysis of the *Researcher* networks was minimised, when compared to the results of the *Topic* networks analysis.

10.3 Improvements

The potential improvements of the project solely concern the analysis carried out on the *Researcher* networks, as additional data can be collected and analysed. By collecting additional data about the researchers such as their department and organisation, would enable the discovery of potential correlations between the way researchers are clustered and their department, organisation or location.

10.4 Future work

Any research project, this included, can be extended to include more data or further experiments or analysis. In this case, four possible ways of extending the work carried out were identified:

- Extending comparison experiments stage to include more community detection algorithms
- Extending the analysis of the *Researcher* networks
- Incorporate further data as node and edge attributes of the networks
- Comparison to a study carried out within a different context e.g. citation networks

Firstly, the comparison experiments stage which is partially concerned with the identification of an optimal community detection algorithm for the data, can be extended to include additional community detection algorithms to the ones already considered. This study carried out experiments using all

community detection algorithms provided by the iGraph network analysis package. However, other packages are available, and consist of algorithms which are not included in iGraph. Therefore, these algorithms could be implemented and Incorporated into the experiments in order to determine whether better clustering results than the ones produced by the *Louvain* community detection algorithm can be achieved.

Secondly, the analysis of the *Researcher* networks was restricted by the amount of time available. By nature, the *Researcher* networks are less informative than the *Topic* networks as topics can be easily identified by people without additional knowledge, while researchers are cannot unless those same people know the researchers or work within EPSRC. However, the *EPSRC GoW* service provides substantial public data, which is partly used in this project, but not fully. Both grant and researcher records consist of additional information which is not used in this project within fields such as the *Department*, *Organisation* and *Industrial Sector Classifications* of a grant or the *Organisation* and *Department* of a researcher. Incorporating additional data in the form of node and edge attributes within both the *Topic* and *Researcher* networks will increase the level of contextual information available and provide a guaranteed extension of the analysis scope.

Finally, another way of extending this work can involve a comparative analysis of the results produced in this project and the results of a study with a similar end goal but which aimed to solve a different kind of practical or scientific problem. For example, this could represent a study involving the analysis of a citation network constructed based on the citations within a significant number of research papers. The objective would be to cluster research papers together in the hope that they share a common subject and researcher topic clusters are revealed. A comparative study may result in valuable insights into potential correlations between the two result sets while, also help cement graph theory as an optimal and valuable solution to clustering problems in both real-world and scientific scenarios.

11 Conclusion

In this project, graph theory was used as a novel approach to a real-world problem, involving the identification of topic and researcher clusters in publicly available data provided by EPSRC. The objective of the project was not only to provide a solution to the problem, but also to determine whether graph theory could provide the solution.

Furthermore, the problem that the project aims to solve was defined and extensive background information about EPSRC, the concept of modularity and community detection algorithms was provided. The state-of-the-art of the related topics was also reviewed. Additionally, the methods put into practice throughout every stage of the project were explained in detail.

The current and historical data collected from EPSRC was interpreted in a number of different ways and lead to several *Topic* and *Researcher* networks being constructed using both the current (2010 to 2016) and historical (1990 to 2000, 2000 to 2010) data sets. This was followed by an extensive comparison experiments on both current and historical data sets which aimed to identify an optimal edge weight and community detection algorithm that would result in a highly accurate and coherent clustering of topics and researchers. The candidates considered included three different interpretations of the edge weight attribute (*unweighted*, *weighted by normalised number of grants*, *weighted by normalised value of grants*) and eight different community detection algorithms including *Louvain*, *Spinglass* and *Fast Greedy*.

The comparative analysis resulted in a significant and valuable set of results. Firstly, edges *weighted by the normalised number of grants* was determined as the optimal interpretation of the edge weight attribute due to its high modularity score and rational clustering produced. Secondly, the *Louvain* method was "crowned" as the optimal community detection algorithm due to its high performance in the experiments and the well-defined nature of the community structure it identified. Finally, using the *Topic-grant* and

Researcher-topic networks proved to be a better option than other network interpretations, as more cogent and balanced clusters of topics and researchers were produced.

This thesis, based on the knowledge available, represents the first approach of deploying graph theory in order to provide a solution to a real-world problem which at the moment, is specific to one organisation - EPSRC. With the data undergoing extensive experimental comparison, an evaluated solution featuring surprisingly high performance is identified and proposed.

In conclusion, this project represents clear evidence that the novel approach based on graph theory is of great value and, because it is not limited by any data set, it can be used to identify solutions to other real-world problems, perhaps the identification of topic communities within a network of newspaper articles.

12 Glossary

This glossary explains how the network naming convention was formulated while also specifying the long form of the abbreviations used in the thesis.

12.1 Naming convention

Throughout this thesis project, networks are mentioned using a standard naming convention, as follows:

term1 dash term2 network

where:

- term1 means that nodes represent term1 in the network
- term2 means that edges represent term2 in the network

The standard naming convention is used to name the 4 networks constructed in this project, as follows:

- **Topic-grant network**, in which:
  - first term (topic) means that nodes represent topics
  - second term (grant) means that edges represent grants
- **Topic-researcher network**, in which:
  - first term (topic) means that nodes represent topics
  - second term (researcher) means that edges represent researchers
- **Researcher-grant network**, in which:
  - first term (researcher) means that nodes represent researchers
  - second term (grant) means that edges represent grants
- **Researcher-topic network**, in which:
  - first term (researcher) means that nodes represent researchers
  - second term (topic) means that edges represent topics

12.2 List of abbreviations

A number of abbreviations are also used in the thesis report, particularly in the comparison experiments, and their long form is the following:

- **EPSRC** stands for **E**ngineering and **P**hysical **S**ciences **R**esearch **C**ouncil
- **GoW** stands for **G**rants **o**n the **W**eb

- **uw** stands for edge **u**n**w**eighted
- **wnn** stands for edge **w**eighted by **n**ormalized **n**umber of grants
- **wnv** stands for edge **w**eighted by **n**ormalized **v**alue of grants

- **SG** stands for **S**ping**g**lass community detection algorithm
- **LV** stands for **L**ou**v**ain community detection algorithm
- **FG** stands for **F**ast **G**reedy community detection algorithm

References

[1] EPSRC. About us. URL: https://www.epsrc.ac.uk/about/ (visited on 28/07/2016).

[2] EPSRC. Our portfolio. URL: https://www.epsrc.ac.uk/research/ourportfolio/ (visited on 29/07/2016).

[3] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[4] ACM. Acm digital library. http://dl.acm.org/.

[5] Yeyun Gong, Qi Zhang, Xuyang Sun, and Xuanjing Huang. Who will you@? In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 533–542. ACM, 2015.

[6] Marina Sokolova, Kanyi Huang, Stan Matwin, Joshua Ramisch, Vera Sazonova, Renee Black, Chris Orwa, Sidney Ochieng, and Nanjira Sambuli. Topic modelling and event identification from twitter textual data. *arXiv preprint arXiv:1608.02519*, 2016.

[7] Chenguang Wang, Yangqiu Song, Dan Roth, Ming Zhang, and Jiawei Han. World knowledge as indirect supervision for document clustering. *arXiv preprint arXiv:1608.00104*, 2016.

[8] Rocco Tripodi and Marcello Pelillo. Document clustering games in static and dynamic scenarios. *arXiv preprint arXiv:1607.02436*, 2016.

[9] Koushiki Sarkar and Ritwika Law. A novel approach to document classification using wordnet. *arXiv preprint arXiv:1510.02755*, 2015.

[10] Gabriela Csurka. Document image classification, with a specific view on applications of patent images. *arXiv preprint arXiv:1601.03295*, 2016.

[11] Vanessa Kitzie and Debanjan Ghosh. #criming and #alive: Network and content analysis of two sides of a story on twitter. *Proceedings of the Association for Information Science and Technology*, 52(1):1–10, 2015.

[12] Hiroyuki Kasai, Wolfgang Kellerer, and Martin Kleinsteuber. Network volume anomaly detection and identification in large-scale networks based on online time-structured traffic tensor tracking. 2016.

[13] Google. Google books ngram viewer. URL: `https://books.google.com/ngrams` (visited on 08/09/2016).

[14] JetBrains. Pycharm. URL: `https://www.jetbrains.com/pycharm/` (visited on 30/07/2016).

[15] Microsoft. Excel. URL: `http://office.microsoft.com/en-us/excel` (visited on 30/07/2016).

[16] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5):1–9, 2006.

[17] A Hagberg, D Schult, and P Swart. Networkx. URL: `https://networkx.github.io/index.html` (visited on 30/07/2016).

[18] Jonathan Feinberg. Wordle. URL: `http://www.wordle.net/` (visited on 30/07/2016).

[19] Adobe. Photoshop. URL: `http://adobe.com/photoshop` (visited on 30/07/2016).

[20] Jörg Reichardt and Stefan Bornholdt. Statistical mechanics of community detection. *Physical Review E*, 74(1):016110, 2006.

[21] Günce Keziban Orman, Vincent Labatut, and Hocine Cherifi. On accuracy of community structure discovery algorithms. *arXiv preprint arXiv:1112.4134*, 2011.

[22] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[23] Mark EJ Newman. Fast algorithm for detecting community structure in networks. *Physical review E*, 69(6):066133, 2004.

A Data Management Plan

A Research Data Management Plan was compiled and can be accessed at the following address:

`https://github.com/SergiuTripon/msc-thesis-na-epsrc/blob/master/documents/research-data-management-plan/pdf/research_data_management_plan.pdf`

B EPSRC grant data

Table B.1: Topics clustered within each sub-community of Community 1 discovered by the optimal solution identified in the experiment on the *Topic-grant* network constructed using the current (2010 to 2016) data set. **C1** stands for Community 1. Rows representing sub-communities are named using second-level labels (1.1, 1.2 and so on).

C1	**1.1**	ageing: chemistry/biochemistry, analytical science, biomedical sciences
	1.2	biomaterials, med.instrument.device& equip., biomechanics & rehabilitation, medical imaging, biomedical neuroscience, novel industrial products, development (biosciences), systems neuroscience, drug formulation & delivery, tissue engineering, mathematical & statistic psych
	1.3	drug formulation & delivery, tissue engineering, mathematical & statistic psych, bioelectronic devices, medical science & disease, bioinformatics, microbiology, cells, population ecology, complex fluids & soft solids, theoretical biology, genomics
	1.4	biological & medicinal chem., protein chemistry, catalysis & enzymology, protein folding / misfolding, chemical biology, structural biology
C2	**2.1**	artificial intelligence, information & knowledge mgmt, behavioural & experimental eco, intelligent measurement sys., comput./corpus linguistics, international law, computational linguistics, marketing, criminal law & criminology, psychology, criminology, science & technology studies, governance, social policy
	2.2	cognitive psychology, image & vision computing, cognitive science appl. in ict, mental health, composition, music & acoustic technology, design processes, musical performance, developmental psychology, new & emerging comp. paradigms, human communication in ict, robotics & autonomy, human-computer interactions, vision & senses - ict appl.
	2.3	macroeconomics, political geography
	2.4	animal behaviour, networks & distributed systems, computer sys. & architecture, organisational studies, data handling & storage, parallel computing, digital signal processing, rf & microwave technology, fundamentals of computing, social psychology, industrial-org/occupational, software engineering, international relations theory, system on chip, knowledge management, vlsi design, modelling & simul. of it sys.

Table B.1: Continued from previous page.

	2.5	applied arts htp, mobile computing, computer graphics & visual., multimedia, design engineering, new media/web-based studies, digital art & design, product design, digital arts htp, social anthropology, manufact. business strategy, social theory, media & communication studies, time-based media htp
C3	**3.1**	acoustics, fluid dynamics, aerodynamics, heat & mass transfer, assess/remediate contamination, microsystems, bioenergy, multiphase flow, coal technology, pollution, combustion, power sys man, prot & control, control engineering, power systems plant, development geography, rheology, earth engineering, separation processes, electric motor & drive systems, underwater engineering, energy - conventional, wind power, energy - marine & hydropower
	3.2	biochemical engineering, macro-molecular delivery, bioprocess engineering, manufact. enterprise ops& mgmt, design of process systems, manufacturing machine & plant, food processing, particle technology, food structure/composition, protein engineering, intelligent & expert systems
	3.3	asymmetric chemistry, gas & solution phase reactions, carbohydrate chemistry, materials characterisation, catalysis & applied catalysis, materials processing, chemical structure, materials synthesis & growth, chemical synthetic methodology, physical organic chemistry, co-ordination chemistry, plant physiology, electrochemical science & eng., plant responses to environment, electromagnetics, reactor engineering, evolution & populations, surfaces & interfaces
	3.4	carbon capture & storage, instrumentation eng. & dev., diamond light source, materials testing & eng., energy storage, mech. & fluid power transmiss., eng. dynamics & tribology, oil & gas extraction, fuel cell technologies
	3.5	research approaches, synthetic biology
C4	**4.1**	mathematical aspects of or, microeconomic theory
	4.2	algebra & geometry, mathematical physics, continuum mechanics, non-linear systems mathematics, logic & combinatorics, numerical analysis, mathematical analysis, statistics & appl. probability
C5	**5.1**	complexity science, management & business studies, economics, social stats., comp. & methods, education, sociology, environmental planning, sustainable energy networks, human geography (general), urban & land management
	5.2	animal organisms, energy - nuclear, climate & climate change, land - ocean interactions, coastal & waterway engineering, regional & extreme weather, earth & environmental

Table B.1: Continued from previous page.

	5.3	building ops & management, pavement engineering, civil engineering materials, structural engineering, construction ops & management, sustainable energy vectors, energy efficiency, waste management, environment & health, waste minimisation, environmental economics, water engineering
	5.4	geohazards, survey & monitoring, ground engineering, transport ops & management, soil science
C6	**6.1**	design & testing technology, optical devices & subsystems, displays, optical phenomena, electronic devices & subsys., optoelect. devices & circuits, lasers & optics, power electronics, optical communications
	6.2	biological membranes, magnetism/magnetic phenomena, biophysics, solar technology, condensed matter physics, tools for the biosciences, high performance computing
	6.3	computational methods & tools, plasmas - laser & fusion, fusion, plasmas - technological
	6.4	atoms & ions, quantum fluids & solids, cold atomic species, quantum optics & information, light-matter interactions, scattering & spectroscopy

Table B.2: Topics clustered within each community and sub-community discovered as a result of applying the Louvain community detection algorithm to the *Topic-grant* network constructed using the historical (2000 to 2010) data set. Six of the row names are abbreviated: **C1** stands for Community 1, **C2** for Community 2 and so on. Rows representing sub-communities of a community are named using second-level labels (1.1, 1.2, 2.1, 2.2 and so on).

C1	**1.1**	analytical science, co-ordination chemistry, asymmetric chemistry, combinatorial chemistry, biological & medicinal chem., electrochemical science & eng., carbohydrate chemistry, mantle & core processes, catalysis & applied catalysis, physical organic chemistry, chemical synthetic methodology, reactor engineering
	1.2	astron. & space sci. technol., light-matter interactions, atoms & ions, magnetism/magnetic phenomena, catalysis & enzymology, nuclear structure, chemical structure, optical phenomena, cold atomic species, plasmas - laser & fusion, condensed matter physics, plasmas - technological, galactic & interstellar astron, quantum fluids & solids, gas & solution phase reactions, quantum optics & information, high performance computing, scattering & spectroscopy, lasers & optics, surfaces & interfaces
	1.3	bioelectronic devices, medical science & disease, electronic devices & subsys., microsystems, instrumentation eng. & dev., musculoskeletal system, materials characterisation, optical communications, materials processing, optical devices & subsystems, materials synthesis & growth, optoelect. devices & circuits
	1.4	biological membranes, chemical biology, bionanoscience, protein chemistry, bionanotechnology, protein folding / misfolding, biophysics, structural biology
	1.5	biomaterials, genomics, bioprocess engineering, particle technology, cells, rheology, complex fluids & soft solids, separation processes, development (biosciences), stem cell biology, drug formulation & delivery, synthetic biology, food processing, tissue engineering, food structure/composition
C2	**2.1**	aerodynamics, manufact. business strategy, control engineering, manufact. enterprise ops& mgmt, design & testing technology, manufacturing machine & plant, electromagnetics

Table B.2: Continued from previous page.

	2.2	algebra & geometry, mathematical aspects of or, animal & human physiology, mathematical physics, complexity science, multiphase flow, continuum mechanics, non-linear systems mathematics, design of process systems, numerical analysis, evolution & populations, population ecology, fluid dynamics, statistics & appl. probability, logic & combinatorics, theoretical biology, mathematical analysis, upper atmos process & geospace
	2.3	acoustics, materials testing & eng., assess/remediate contamination, mech. & fluid power transmiss., building ops & management, pavement engineering, civil engineering materials, structural engineering, coastal & waterway engineering, transport ops & management, construction ops & management, urban & land management, design engineering, waste management, eng. dynamics & tribology, waste minimisation, ground engineering, water engineering
C3	**3.1**	bioenergy, fuel cell technologies, carbon capture & storage, fusion, electric motor & drive systems, power sys man, prot & control, energy - conventional, solar technology, energy - nuclear, sustainable energy networks, energy efficiency, sustainable energy vectors, energy storage, wind power
	3.2	microbiology, power electronics
	3.3	coal technology, mining & minerals extraction, combustion, safety & reliability of plant, heat & mass transfer
	3.4	energy - marine & hydropower, power systems plant, oil & gas extraction, underwater engineering
C4	**4.1**	crop science, soil science
C5	**5.1**	artificial intelligence, languages & linguistics, bioinformatics, modelling & simul. of it sys., biomechanics & rehabilitation, new & emerging comp. paradigms, biomedical neuroscience, parallel computing, cognitive science appl. in ict, rf & microwave technology, computer sys. & architecture, robotics & autonomy, digital signal processing, software engineering, fundamentals of computing, system on chip, image & vision computing, vision & senses - ict appl., intelligent measurement sys., vlsi design

Table B.2: Continued from previous page.

	5.2	applied arts htp, media & communication studies, cultural history, mental health, design htp, mobile computing, design processes, multimedia, digital art & design, music & acoustic technology, digital arts htp, networks & distributed systems, economic & social history, new media/web-based studies, economics, policy, arts mgmt & creat ind, intelligent & expert systems, product design, language acquisition, publishing, language training/educational, social stats., comp. & methods, management & business studies, sociology, med.instrument.device& equip.
	5.3	cultural studies & pop culture, pollution, displays, psychology, information & knowledge mgmt
	5.4	accelerator r&d, escience, agricultural systems, human communication in ict, applied linguistics, human geography, archaeology of literate soc., human-computer interactions, comput./corpus linguistics, interpreting & translation, computer graphics & visual., medical imaging, drama & theatre - other, psycholinguistics, education, science-based archaeology, environmental informatics, sociolinguistics, environmental planning

Table B.3: Topics clustered within each community and sub-community discovered as a result of applying the Louvain community detection algorithm to the *Topic-grant* network constructed using the historical (1990 to 2000) data set. Six of the row names are abbreviated: **C1** stands for Community 1, **C2** for Community 2 and so on. Rows representing sub-communities of a community are named using second-level labels (1.1, 1.2, 2.1, 2.2 and so on).

C1	**1.1**	acoustics, intelligent measurement sys., aerodynamics
	1.2	coal technology, energy - conventional, combustion, safety & reliability of plant
	1.3	bioprocess engineering, heat & mass transfer, cells, multiphase flow, complex fluids & soft solids, particle technology, design of process systems, reactor engineering, fluid dynamics, rheology
	1.4	building ops & management, transport ops & management, computer graphics & visual., urban & land management, energy efficiency, wind power, intelligent & expert systems
	1.5	bioenergy, waste minimisation, waste management, water engineering
C2	**2.1**	condensed matter physics, materials synthesis & growth, magnetism/magnetic phenomena, quantum fluids & solids, materials characterisation, solar technology, materials processing, sustainable energy networks
	2.2	displays, optoelect. devices & circuits, electronic devices & subsys., power electronics, microsystems, system on chip, optical communications, vlsi design, optical devices & subsystems
	2.3	atoms & ions, optical phenomena, cold atomic species, plasmas - laser & fusion, lasers & optics, quantum optics & information, light-matter interactions, scattering & spectroscopy
C3	**3.1**	development (biosciences), modelling & simul. of it sys., drug formulation & delivery, theoretical biology
	3.2	algebra & geometry, mathematical analysis, fundamentals of computing, mathematical physics
	3.3	continuum mechanics, numerical analysis, logic & combinatorics, parallel computing, mathematical aspects of or, population ecology, medical science & disease, statistics & appl. probability, non-linear systems mathematics
C4	**4.1**	control engineering, manufacturing machine & plant, electric motor & drive systems, mech. & fluid power transmiss.

Table B.3: Continued from previous page.

	4.2	bioelectronic devices, mobile computing, cognitive science appl. in ict, multimedia, digital signal processing, networks & distributed systems, electromagnetics, rf & microwave technology, human communication in ict, vision & senses - ict appl., human-computer interactions
	4.3	assess/remediate contamination, ground engineering, civil engineering materials, materials testing & eng., coastal & waterway engineering, music & acoustic technology, energy - nuclear, oil & gas extraction, eng. dynamics & tribology, pavement engineering
	4.4	construction ops & management, manufact. business strategy, design & testing technology, manufact. enterprise ops& mgmt, design engineering, nuclear structure, design processes, software engineering, information & knowledge mgmt
	4.5	artificial intelligence, tools for the biosciences, image & vision computing, underwater engineering,robotics & autonomy
C5	**5.1**	biological & medicinal chem., chemical synthetic methodology, catalysis & enzymology, co-ordination chemistry, chemical biology, gas & solution phase reactions, chemical structure, physical organic chemistry
	5.2	asymmetric chemistry, catalysis & applied catalysis, carbohydrate chemistry, combinatorial chemistry
	5.3	analytical science, plasmas - technological, instrumentation eng. & dev., surfaces & interfaces
	5.4	biomaterials, power systems plant, med.instrument.device& equip., tissue engineering, power sys man, prot & control
	5.5	electrochemical science & eng., mining & minerals extraction, energy storage, separation processes, fuel cell technologies, sustainable energy vectors

Table B.4: Topics clustered within each community and sub-community discovered as a result of applying the Louvain community detection algorithm to the *Topic-researcher* network constructed using the current (2010 to 2016) data set. Six of the row names are abbreviated: **C1** stands for Community 1, **C2** for Community 2 and so on. Rows representing sub-communities of a community are named using second-level labels (1.1, 1.2, 2.1, 2.2 and so on).

C1	**1.1**	building ops & management, manufact. business strategy, civil engineering materials, pavement engineering, construction ops & management, robotics & autonomy, design & testing technology, soil science, earth engineering, structural engineering, environmental economics, survey & monitoring, geohazards, waste management, ground engineering, wind power
	1.2	animal organisms, environment & health, climate & climate change, food processing, coastal & waterway engineering, land - ocean interactions, earth & environmental, water engineering, energy - marine & hydropower
	1.3	aerodynamics, intelligent measurement sys., complexity science, management & business studies, education, social stats., comp. & methods, energy efficiency, sociology, environmental planning, sustainable energy networks, human geography (general), transport ops & management, intelligent & expert systems, urban & land management
C2	**2.1**	cognitive psychology, information & knowledge mgmt, comput./corpus linguistics, marketing, computational linguistics, mental health, design processes, mobile computing, developmental psychology, psychology, human communication in ict, social theory, human-computer interactions
	2.2	behavioural & experimental eco, mathematical & statistic psych, criminal law & criminology, media & communication studies, criminology, modelling & simul. of it sys., governance, organisational studies, international law, social policy, knowledge management, social psychology
	2.3	bioinformatics, design engineering, biomedical neuroscience, new & emerging comp. paradigms, cognitive science appl. in ict, research approaches, control engineering, vision & senses - ict appl.
	2.4	computer sys. & architecture, parallel computing, data handling & storage, political geography, fundamentals of computing, product design, industrial-org/occupational, software engineering, international relations theory, underwater engineering, networks & distributed systems, vlsi design

Table B.4: Continued from previous page.

	2.5	animal behaviour, macroeconomics, applied arts htp, multimedia, artificial intelligence, music & acoustic technology, composition, musical performance, computer graphics & visual., new media/web-based studies, digital art & design, science & technology studies, digital arts htp, social anthropology, digital signal processing, time-based media htp, image & vision computing
C3	**3.1**	genomics, theoretical biology
	2.2	algebra & geometry, non-linear systems mathematics, continuum mechanics, numerical analysis, logic & combinatorics, regional & extreme weather, mathematical analysis, statistics & appl. probability, mathematical physics
	2.3	acoustics, microeconomic theory, mathematical aspects of or, rheology
C4	**4.1**	ageing: chemistry/biochemistry, magnetism/magnetic phenomena, analytical science, materials characterisation, atoms & ions, materials synthesis & growth, bioelectronic devices, optical communications, biological membranes, optical devices & subsystems, biomedical sciences, optical phenomena, bionanoscience, optoelect. devices & circuits, biophysics, plant physiology, carbohydrate chemistry, plant responses to environment, chemical structure, pollution, cold atomic species, quantum fluids & solids, complex fluids & soft solids, quantum optics & information, condensed matter physics, rf & microwave technology, displays, scattering & spectroscopy, electronic devices & subsys., solar technology, evolution & populations, surfaces & interfaces, high performance computing, system on chip, lasers & optics, tools for the biosciences, light-matter interactions
	4.2	accelerator r&d, fuel cell technologies, assess/remediate contamination, fusion, bioenergy, instrumentation eng. & dev., carbon capture & storage, manufact. enterprise ops& mgmt, coal technology, manufacturing machine & plant, combustion, materials processing, computational methods & tools, materials testing & eng., development geography, mech. & fluid power transmiss., economics, oil & gas extraction, electric motor & drive systems, plasmas - laser & fusion, electromagnetics, plasmas - technological, energy - conventional, power sys man, prot & control, energy - nuclear, power systems plant, energy storage, sustainable energy vectors, eng. dynamics & tribology, waste minimisation, food structure/composition

Table B.4: Continued from previous page.

	4.3	asymmetric chemistry, gas & solution phase reactions, biochemical engineering, heat & mass transfer, bioprocess engineering, macro-molecular delivery, catalysis & applied catalysis, microsystems, chemical biology, multiphase flow, chemical synthetic methodology, particle technology, co-ordination chemistry, physical organic chemistry, design of process systems, protein engineering, diamond light source, reactor engineering, electrochemical science & eng., separation processes, fluid dynamics, synthetic biology
	4.4	biological & medicinal chem., microbiology, biomaterials, novel industrial products, biomechanics & rehabilitation, population ecology, catalysis & enzymology, power electronics, cells, protein chemistry, development (biosciences), protein folding / misfolding, drug formulation & delivery, structural biology, med.instrument.device& equip., systems neuroscience, medical imaging, tissue engineering, medical science & disease

C Source Code

During this project, source code was written in Python and Bash and stored within **12 files (6,123 lines)**, as follows:

1. **Data collection (2,541 lines)**

    1.1. **download.sh (153 lines)**, written to download grant and research records from the *EPSRC GoW service*

    1.2. **extract.py (964 lines)**, written to extract information from the downloaded data

    1.3. **link.py (844 lines)**, written to establish the links between the topics and researchers

    1.4. **make.py (414 lines)**, written to create the *Topic* and *Researcher networks*

    1.5. **run.py (35 lines)**, written to run the *extract.py*, *link.py*, *make.py* functions from one central file

    1.6. **check.sh (131 lines)**, written to ensure correct file creation

2. **Network analysis (3,582 lines)**

    2.1. **network.py (711 lines)**, written to analyse the network

    2.2. **communities.py (598 lines)**, written to analyse the communities within the network

    2.3. **sub_communities.py (423 lines)**, written to analyse the sub-communities in each community within the network

    2.4. **analysis.py (1,176 lines)**, written to bind the *network.py*, *communities.py*, *sub_communities.py* functions together in one central file

    2.5. **evaluation.py (173 lines)** written to evaluate the community structure identified

    2.6. **clean.sh (501 lines)**, written to delete multiple files from specific folders

C.1 Source code location

The source code is stored in a GitHub repository located at the following web address:

`https://github.com/SergiuTripon/msc-thesis-na-epsrc`

C.2 Running the source code

Note: In order to run the source code, an virtual environment installation is required. The code is written in Python 3.5. The packages used in the project are listed in the *requirements.txt* file and can be install using *pip*.

Running the network analysis is achieved by running the *analysis.py* file with the desired parameters (**-n** requires network (topic or researcher), **-i** requires interpretation (grants, researchers or topics), **-d** requires data set (1990-2000, 2000-2010, 2010-2016)), following the steps below:

```
# activate virtual environment
$ source venv/bin/activate

# navigate to analysis source folder
$ cd msc-thesis-na-epsrc/analysis/src/

# analyse topic (grants as edges, 2010-2016)
$ python analysis.py -n topic -i grants -d 2010-2016
# analyse topic (grants as edges, 2000-2010)
$ python analysis.py -n topic -i grants -d 2000-2010
# analyse topic (grants as edges, 1990-2000)
$ python analysis.py -n topic -i grants -d 1990-2000

# analyse topic (researchers as edges, 2010-2016)
$ python analysis.py -n topic -i researchers -d 2010-2016
# analyse topic (researchers as edges, 2000-2010)
$ python analysis.py -n topic -i researchers -d 2000-2010
# analyse topic (researchers as edges, 1990-2000)
```

```
$ python analysis.py -n topic -i researchers -d 1990-2000

# analyse researcher (grants as edges, 2010-2016)
$ python analysis.py -n topic -i grants -d 2010-2016
# analyse researcher (researchers as edges, 2000-2010)
$ python analysis.py -n topic -i grants -d 2000-2010
# analyse researcher (researchers as edges, 1990-2000)
$ python analysis.py -n topic -i grants -d 1990-2000

# analyse researcher (topics as edges, 2010-2016)
$ python analysis.py -n topic -i topics -d 2010-2016
# analyse researcher (topics as edges, 2000-2010)
$ python analysis.py -n topic -i topics -d 2000-2010
# analyse researcher (topics as edges, 1990-2000)
$ python analysis.py -n topic -i topics -d 1990-2000
```

C.3 GitHub Wiki

A GitHub Wiki was created to accompany this project and is located at the following web address:

`https://github.com/SergiuTripon/msc-thesis-na-epsrc/wiki`

Note: The information on the GitHub Wiki may be out-of-date.

C.4 Code snippets

This section presents two code snippets which are considered particularly important. The first one achieves the contrast between grants as edges and grant records, while the second normalises the values of the node and edge attributes.

C.4.1 Contrast between grants as edges and grant records

This snippet of code was written in order to convert edges within the network in grants. Moreover, it calculates the number and value of grants within communities, between communities and within the entire network. Achieving this was extremely beneficial as it enabled the analysis of research trend and funding on the current (2010 to 2016) and historical (1990 to 2000, 2000 to 2010) data sets.

Listing C.1: Code snippet showing function written to turn edges into grants and calculate the number and value of grants within network and communities and between communities.

```
# turns edges into grants
def turn_edges_into_grants(community, edge_type, method, count1,
    path):

    # variable to split path
    path_split = path.split('/', 3)

    # variable to hold temporary path
    path_temp = ''

    # if split path equals to current
    if path_split[1] == 'current':
        # set temporary path
        path_temp = '{}'.format(path_split[1])
    # if split path equals to past
    elif path_split[1] == 'past':
        # set temporary path
        path_temp = '{}/{}'.format(path_split[1], path_split[2])

    # variable to hold input file
    input_file = open(r'../../network-maker/output/grants/'
```

```
'{}/info/grant_{}.pkl'.format(path_temp, path_split[0]), 'rb')
# load data structure from file
grant_entities = load(input_file)
# close input file
input_file.close()

# variable to hold entity links
entity_links = [[community.vs['label'][edge.source],
    community.vs['label'][edge.target]]
                for edge in community.es()]

# variable to hold grants
grants = OrderedDict()

# if split path equals to topics
if path_split[0] == 'topics':

    # set grants
    grants = OrderedDict((ref, attr[1]) for entity_link in
        entity_links for ref, attr in grant_entities.items()
                        if entity_link[0] and entity_link[1] in
                            attr[0])

# if split path equals to researchers
elif path_split[0] == 'researchers':

    # set grants
    grants = OrderedDict((ref, attr[1]) for entity_link in
        entity_links for ref, attr in grant_entities.items()
                        if entity_link[0] and entity_link[1] in
                            [researcher[0] for researcher in
                            attr[0]])
```

```
# variable to hold number
number = len([ref for ref in grants.keys()])

# set locale to Great Britain
setlocale(LC_ALL, 'en_GB.utf8')

# variable to hold value
value = sum([attr for attr in grants.values()])

# variable to hold output file
output_file =
    open('../../data/networks/{}/communities/txt/{}/{}/'
                 'grants.txt'.format(path, edge_type, method),
                    mode='a')

# if count1 is equal to 1
if count1 == 1:

    # write header to file
    output_file.write('> Number and value of grants in each
        community\n\n')

    # write grant number and value to file
    output_file.write('- Community {}: {:>4d}
        {}\n'.format(count1, number, currency(value,
        grouping=True)))

# if count1 is not equal to 1
else:

    # write grant number and value to file
```

```
        output_file.write('- Community {}: {:>4d}
            {}\n'.format(count1, number, currency(value,
            grouping=True)))

    # return number and value
    return grants, number, value
```

C.4.2 Normalisation of node and edge attribute values

This code snippet was written in order to normalise the values of the node and edge attributes. The value of grants attribute, in particular, consisted of very large values, up to 7 digits. By normalising the values of the node and edge attributes, working with such large values was avoided and, therefore, the development, analysis and visualisation process was improved.

Listing C.2: Code snippet showing function written to normalise the values of the node and edge attributes.

```
# normalises values
def norm_vals(vals, new_min, new_max):

    # variable to hold old minimum and maximum
    old_min, old_max = min(vals), max(vals)
    # variable to hold old and new range
    old_range, new_range = old_max - old_min, new_max - new_min

    int_vals = [int(val) for val in vals]

    # variable to hold new values
    new_vals = [round((((val - old_min) * new_range / old_range) +
        new_min), 0) for val in int_vals]

    # return new values
    return new_vals
```

D Supplementary material

Additionally to the appendix in the thesis report, a 1095-page document of supplementary material was also produced and can be accessed either in the submission or at the following web address:

`https://github.com/SergiuTripon/msc-thesis-na-epsrc/blob/master/documents/supplementary-material/15110029_sergiu_tripon_supplementary_material.pdf`