Coherent clustering of topics and researchers in networks based on EPSRC data: a novel approach to a real-world problem

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 3, 2016

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 in 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 normalized 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 genre, type and language. Their user account also provides useful information such as gender, age, when and how often they use Netflix, the platform 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 into the data. The results of this research project can then have an impact in decision-making. For example, when deciding which series Netflix should add next to the platform based on user demand of a specific genre. 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 missing.

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 fine or coarse 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 research communities. The project aims to achieve this through applying a novel approach to a real world problem, involving graph theory, network science, the concept of modularity and community detection. Therefore, as much as the objectives are the clustering of topics and researchers, verifying whether this approach can be used to

solve this kind of problems is also a crucial objective. It is expected that this approach will provide a rational 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, five 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 network (Grants as edges)* produced a more reasonable topic clustering in comparison to the *Researchers as edges* interpretation of it. In contrast, the *Researcher network (Topics as edges)* network performed better than the *Researcher network (Grants as edges)* when considering the researcher clustering. Furthermore, the edge weight normalized by the 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 the previous chapter, the project was introduced and the problem was defined. This chapter provides detailed background information about EPSRC and the problem that this project aims to solve using several concepts including graph theory, modularity and community structure and detection, which are also presented. Furthermore, other fields related to this study are also briefly introduced.

2.1 EPSRC

Engineering and Physical Sciences Research Council (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 manage and fund research projects in various disciplines. It ensures effective collaborative work between the councils enhancing the overall impact of their research, training, and innovation.

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 3175 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 3175 grants are classified using 225 current topics. Researchers submit funding proposals to EPSRC, which once accepted turn into research grants financially supported by EPSRC.

During the process of making a proposal, one of the tasks that re-

searchers complete is the classification of the project using a number of different topics. They are allowed to classify the project using one or more topics. By doing this, researchers use their judgement and provide intelligence which can prove to be extremely useful, when analysing a network of topics constructed based on the topic classification of grants. It is essential to note that a community detection method solely focuses on a network and its structure and does not use any other information. However, the structure of that network represents the human judgement and intelligence provided by researchers at the very beginning, during the process of making a proposal.

Assume a problem concerning the way topics within EPSRC are defined and categorised. It is not known how fine or coarse 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 available to cover the full topic classification spectrum of the grants. The relation between topics is also unknown. One approach to solving such a problem is using 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. Visually, imagine that the topics are represented by circles and the links between topics by lines.

Initially, within the network of topics constructed, there is no difference between topics, they are all treated and size the same. The only difference is in 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, a potential alternative could involve grouping similar topics manually, and subsequently analysing each group separately. This would lower the difficulty level compared to the previous attempt, but will significantly increase the time it would take to complete the task. However, if a technique employed computationally would be employed to divide the network into a number of parts, the task would be com-

pleted efficiently and with ease. Using the same computational technique, each part could be divided further into a number of sub-parts in order to obtain a more refined clustering. Additionally, this would allow the analysis of specific parts of the network which will help decision-making regarding the definition of topics tremendously. Furthermore, this solution could also be used to analyse other networks like a network of researchers where the task would involve identifying communities of researchers.

This project aims to divide networks constructed from EPSRC data into clusters of topics representing researcher areas and clusters of researchers using graph theory and community detection methods. Achieving this will bring clarity, efficiency and substantial value to the process of defining topics within EPSRC.

2.2 Network Community Structure

Detecting the community structure in a network is one of several tasks in network science. Usually, the community structure is detected through the use of a community detection algorithm. Over the years, the task has become increasingly popular which led to the birth of a large number of community detection algorithms.

A community detection algorithm divides a network into a number of clusters, which may be overlapping. If the nodes within each cluster are densely connected, it means that the network holds a community structure. In the case of no overlapping clusters, it means that 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 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 al-

gorithm. 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 deep 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 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. However, if the network is large, this 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.

2.4 Network Community Detection Algorithms

Community detection is a popular subject in the study of networks and this is accurately represented in the significant number of community detection

algorithms that have been and are still developed. This project initially considered and compared eight community detection algorithms including *Infomap* and *Walktrap*. Following an extensive comparative analysis, the number of algorithms was reduced to a final three: *Louvain*, *Spinglass* and *Fast Greedy*. The results identified the *Louvain* method as the optimal community detection algorithm. This section provides further information regarding the final three community detection algorithms.

2.4.1 Louvain introduced by Blondel et al.

Louvain is a modularity optimisation algorithm introduced by *Blondel et al.* [4]. It proposes a two-phase hierarchical agglomerative approach which is an improvement of Fast Greedy by Newman et al. [5] 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 [6].

2.4.2 Spinglass introduced by Reichardt and Bornholdt

Spinglass is another modularity optimisation proposed by *Reichardt and Bornholdt* [7] which is based on a combination between a popular statistical mechanic model called Potts spin glass, and community structure. The algorithm applies the technique of simulated annealing on Potts in order to achieve an optimal modularity [6].

2.4.3 Fast Greedy introduced by Newman et al.

FastGreedy, an algorithm developed by *Newman et al.* [5] is based on a greedy optimization method 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 algorithm's hierarchical nature, it produces a hierarchy of community structures. The comparison of modularity values determines the best community structure [6].

2.5 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. A classification system is a method of realizing classification. 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.

2.5.1 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 a book's address. 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.

This form of 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, 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 computationally 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 manually carried out by a human being.

2.5.2 Document Classification

Document classification is a classification problem in the fields of library science, information science 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. As previously mentioned, this can be considered the equivalent of researchers using rationale and manually classifying funding proposals using a number of different topics.

In computer science, document classification is accomplished computationally through the use of various document classification algorithms. This application has similarities to the application of community detection as both are based on algorithms and achieved computationally. However, differences exists 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.

2.6 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.

Document clustering and community detection have clear similarities in terms of the resulting rational clustering. In both methods, the members of a clustering hold a strong relation. 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 and lemmatization, removal of stop words and punctuation and the computation of term frequencies. On the other hand, a community detection algorithm is applied to a network and results in a clustering also known as a community structure, which is based on the structure and dynamics of the network.

2.7 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.

Similarly to document clustering, topic modelling 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.

Additionally, 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.

3 Literature Survey

The previous chapter introduced EPSRC and the problem addressed in this research project while also providing background to several concepts in network science, community detection algorithms and other forms of classification. This section presents the literature reviewed prior to commencing this project identify related works and determine whether studies with a similar focus as this one have already been completed.

Community Structure on Wikipedia [8] documents community structure thoroughly and provided a good starting point to the literature survey. The article describes the properties and application of community structures in networks. Moreover, it gives a detailed description of a significant number of community detection algorithms as well as several testing methods enabling the evaluation of the algorithms. The information on Wikipedia helped to gain a general understanding of the subject and identify the further direction of the survey.

Community Detection and Mining in Social Media [9] is a class taught by *Lei Tang* (Yahoo! Labs) and *Huan Liu* (Arizona State University) at Arizona State University. The class follows the teachings of a book with the same name [10] published by *Lei Tang* and *Huan Liu* in 2010. The information from the class is available online in the form of PowerPoint presentations. Essentially, they provide a summary, highlighting the most important parts from the book. The class added the academic knowledge required to the literature survey while also expanding on the information gathered from Wikipedia in regards to community detection evaluation.

Community structure in social and biological networks [11] is a research paper authored by *Michelle Girvan* and *Mark EJ Newman* and published in 2002. The research paper introduces the community structure property, present in many networks. It defines the community structure of

a network as network nodes linked together in densely connected clusters, between which there are only sparser connections.

Girvan and Newman also propose a method to detect communities in networks which is built based on the idea of centrality indices. The method is tested on both *real-world* and *computer-generated* networks whose community structure is both known and unknown. *Real-world* networks used for testing include *Zachary's Karate Club Study* and *College Football*. The study used networks without a known community structure such as the *Collaboration* and *Food Web* networks. In the case of a known community structure, the authors found that the method performed well and identified the known community structure with high-sensitivity and reliability. Furthermore, in both cases of unknown community structure, the method still detected significant and informative community divisions.

This research paper was extremely beneficial as it provided further knowledge in the field of community detection as well as the concepts and methods required to complete this research project.

Finding and evaluating community structure in networks [12] is another paper published in 2004 by *Mark EJ Newman* and *Michelle Girvan* and similar to their previous collaborative work, *Community structure in social and biological networks*. In this publication, the authors propose a set of algorithms for identifying community structure in networks, which is defined as "natural divisions of network nodes into densely connected subgroups." Moreover, this paper represents the introduction of the community structure measure known as *modularity*.

The proposed algorithms include shortest-path betweenness, resistor networks and random walks. A way to measure the strength of the community structure identified by the algorithms proposed is also introduced. The measure provides an objective metric for determining the number of communities a network should be divided into. Testing the algorithms on both *computer-generated* and *real-world networks*, *Newman* and *Girvan* demon-

strate that the algorithms are extremely effective at identifying community structure.

The knowledge gained from this study was extremely helpful throughout the duration of this project. It explained how the algorithms proposed were built and how they work, while also providing examples of networks on which the algorithms were evaluated on. It also showed the progress of *Newman* and *Girvan* in the subject since their previous collaborative research effort.

Topic oriented community detection of rating based social networks is a study conducted by *Reihanian et al.* [13] in 2015 focusing on community detection from the perspective of content analysis. Most community detection research chooses to focus only on the topological structure of the network. In a social network, for example, this is usually based on the number of communications among individuals. In contrast, this research paper aims to go further and explore and analyse the network's content flow.

The development process of the project commences by preprocessing and annotating topic labels and continues with the clustering of social objects and the creation of topic clusters. It concludes by applying a community detection algorithm to the produced topical clusters in order to identify the community structure within each cluster. Furthermore, a number of experiments are carried out on several data sets including *Movielens 100k*, *Book-Crossing*, *CIAO*, *MovieTweetings* and *Movielens Latest*. It makes use of a performance metric, *purity*, as defined by Zhao et al. [14] which considers both topic and linkage structure. It identifies a maximum *purity* value in each experiment as the topical clusters created in each data set incorporate members which are interested in the same unique topics.

Moreover, the study also compares the topic-oriented community detection proposed with the classical community detection method in which topical content is not analysed. It finds higher values of *modularity* and *purity* in the topic-oriented framework, as the basic network is partitioned into top-

ical clusters, and members who have the same topic of interest are clustered into the same identified community.

Topic oriented community detection of rating based social networks has similarities to this project in terms of the focus on topic analysis. It added a new, different perspective to the process of community structure detection which is extremely interesting and definitely a potential path of extending research.

Community detection algorithms: a comparative analysis is a research paper published by *Lancichinetti et al.* [15] in 2009 focusing on the comparison of a wide range of community detection algorithms. Two evaluation benchmarks are employed, the *GN benchmark* by *Girvan* and *Newman* and the *LFR benchmark* proposed by *Lancichinetti et al.*

The community detection algorithms are tested on each evaluation benchmark and include the *Fast greedy modularity optimization*, *Exhaustive modularity optimization via simulated annealing*, *Cfinder*, *Markov Cluster*, *Expectation-maximization* and *Potts model approach*. Furthermore, a number of different graphs were used in the evaluation such as *undirected* and *unweighted* graphs, *directed* and *unweighted* graphs, *undirected* and *weighted* graphs and *undirected* and *unweighted* graphs with overlapping communities. On both evaluation benchmarks, the study found that the *Dynamic algorithm (Infomap)* by *Rosvall* and *Bergstrom* performed the best. The *Fast modularity optimization* by *Blondel et al.* and the *Potts model approach* by *Ronhovde* and *Nussinov* also had a good performance in the evaluation.

This comparative analysis served as a significant source of knowledge in terms of the community detection algorithms available, how they work, when they work best and on which networks. It definitely had an impact on the decisions made in this project in regards to community detection methods utilised.

On Accuracy of Community Structure Discovery Algorithms is another comparative study authored by *Orman et al.* [6] in 2011. It evaluates the

majority of algorithms evaluated in *Community detection algorithms: a comparative analysis*, with the exception of *SpinGlass* by *Reichardt* and *Bornholdt* and *Walktrap* by *Pons* and *Latapy*. A generated benchmark graph using the *LFR benchmark* is used. This means that only artificial networks are taken into consideration while the community structure is already known. Each of the eleven community detection algorithms presented are tested on all generated network samples.

The study found that in all cases the *Dynamic algorithm (Infomap)* by *Rosvall* and *Bergstrom* performed better than all other algorithms. *Infomap* succeeded in identifying the communities even for high mixing coefficient values. Furthermore, *Walktrap*, *Markov Cluster*, *Spinglass* and *Louvain* also had an excellent performance level, although not as good as *Infomap*. The research also discovered that for all algorithms, the higher the degree, the better the performance. Moreover, when the network size increases, some algorithms (*Infomap*, *Infomod*, *Louvain*) performed better, others performed worse (*Commfind*, *SpinGlass*, *LeadingEigenvector*, *Radetal*) while the performance of the remaining algorithms (*Walktrap*, *FastGreedy*, *MarkovCluster*) did not change.

This publication served as a reliability test which determined whether the findings in *Community detection algorithms: a comparative analysis* were consistent when compared to other studies. The findings were indeed consistent as both studies identified more or less the same high-performance and low-performance algorithms. This helped to further solidify the decisions taken surrounding community detection algorithms in this thesis project.

Analysis of Citation Networks is a university project lead by *Anita Valmarska* and *Janez Demšar* at *Jožef Stefan Institute* in *Ljubljana*, *Slovenia*. It focuses on the analysis of citation networks defined as directed networks where one research paper cites another. The data comprises of a collection of 63826 unique psychology-related papers crawled from *Wikipedia* and *Mi-*

crosoft Academic Research data (MAS).

The resulting network consists of 3918 vertices connected by 5732 edges. The study employs the *Louvain* method for the detection of community structure in the created network. The community detection algorithm detected 52 communities with the smallest cluster consisting of 7 research papers, while the largest cluster was constructed of 230 psychology-related publications. The study finds that the results produced have potential as the representation of the communities reveals sensible relationships between psychology sub-fields.

Analysis of Citation Networks is the closest in terms of scale to this research project and provides an example of the data, algorithms and tools that other researchers used. Furthermore, it yields new knowledge and inspiration which contributed to the analysis process of this thesis project.

4 Methodology

The Background section provided information about several techniques and concepts that were used throughout the project. This plays an important role in the development and analysis process of the project which can be divided into the following phases:

1. Collection of data from EPSRC
2. Generation of networks from EPSRC data
3. Formulation of node and edge attributes
    3.1. Normalization of node and edge attribute values
4. Testing of edge weights and algorithms
5. Identification of an optimal edge weight and community detection algorithm
6. Clustering of Topics and Researchers
7. Evaluation of of Topic and Researcher clusters

This section provides background information on the data provided by EPSRC and details each development and analysis phase, describing the methodology followed in the journey of accomplishing its objective.

4.1 Data provided by EPSRC

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

4.1.1 EPSRC Grants on the Web (GoW) service

The Grants on the Web (GoW) service is a EPSRC facility providing information about research grants funded by EPSRC. The service is updated frequently, and consists of large amounts of information regarding current and

historical grants, researchers, panels, quarterly summaries. It also includes search functionality allowing users to search the Web database.

4.1.2 EPSRC Current and Past Grant Portfolio

The Current and Past Grant Portfolio are sub-facilities of the Grants on the Web (GoW) service providing access to current and historical data. Both facilities provide the same type of information, however the access to information differs slightly. The Past Grant Portfolio requires a time period period to be supplied, and provides grant information based on it. This can be a start date, an end date or a start and end date. Fig. 4.1 and Fig. 4.2 present the hierarchical structure of the EPSRC Current and Past Grant Portfolio, respectively.

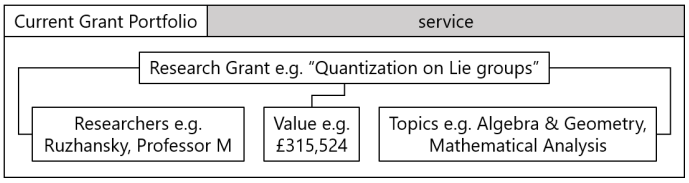


Figure 4.1: Hierarchical structure of the EPSRC Current Grant Portfolio.

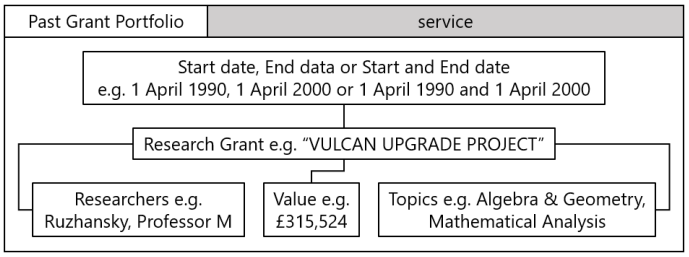
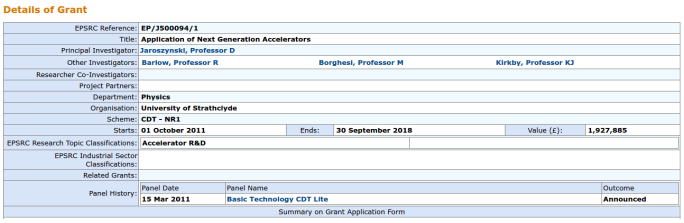


Figure 4.2: Hierarchical Structure of the EPSRC Past Grant Portfolio.

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, topic and industrial sector classification. Fig. 4.3 shows an example of a grant record.

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, organ-

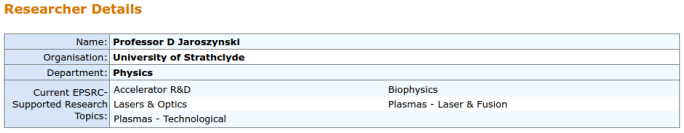


Details of Grant

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 4.3: Grant record within the EPSRC Grants on the Web (GoW) service.

isation, department, current topics, current grants etc. Fig. 4.4 shows an example of a researcher record.



Researcher Details

Name:	Professor D Jaroszynski	
Organisation:	University of Strathclyde	
Department:	Physics	
Current EPSRC-Supported Research Topics:	Accelerator R&D	Biophysics
	Lasers & Optics	Plasmas - Laser & Fusion
	Plasmas - Technological	

Figure 4.4: Researcher record within the EPSRC Grants on Web (GoW) service.

4.2 Collection of data from EPSRC

The previous sections provided description and explanation regarding 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 Grants on the Web (GoW) facility. In terms of the Current and Past Grant Portfolios, data is only extracted from grant and researcher records. In the case of Grant records references, investigators, values and topic classifications are extracted. This project solely extracts the name and current topics from each researcher record. Table 4.1 presents the number of current EPSRC grant and researcher records and historical grant records from which data was collected.

Furthermore, this project uses both current and historical data collected from EPSRC, and 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 grants (post 2010 start date) and

researchers collected on 8 July 2016

- **Historical data set**, consisting of grant records from two time periods:

  - **Ranging from 1990 to 2000**, consisting of grant records which started and ended between 1 April 1990 and 1 April 2000
  - **Ranging from 2000 to 2010**, consisting of grant records which started and ended between 1 April 2000 and 1 April 2010

Table 4.1: 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

The previous section mentioned that both the grant and researcher records are stored as separate web pages within the Grants on the Web (GoW) service.

In order to extract content from a web page, a concept called *web scraping* is used. 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 web page and extracts the text found under specific fields such as reference, investigators, value and topic classifications. Once extracted, the data is validated and then formatted into a format that allows easy manipulation of it. Finally, it is stored within numerous comma-separated files categorised by content and data set.

4.3 Generation of networks from EPSRC data

In the previous section, the data collection process was explained. The data extracted from the Grants on the Web service (GoW) used to construct several networks of topics and researchers structured as follows.

1. **Networks of Topics, with nodes representing Topics**
    1.1. and edges representing Grants, named Grants as edges;
        1.1.1. created using the Current and Historical data sets (1990-2010)
    1.2. and edges representing Researchers, named Researchers as edges;
        1.2.1. constructed using the Current data set only[1]
2. **Networks of Researchers, with nodes representing Researchers**
    2.1. and edges representing Grants, named Grants as edges;
        2.1.1. created using the Current and Historical data sets (1990-2010)
    2.2. and edges representing Topics, named Topics as edges;
        2.2.1. constructed using the Current data set only[1]

Node and edge attributes were also formulated and incorporated into the networks. The attribute values "suffered" a normalization process which decreased the scale of the values. This was followed by extensive experiments which aimed to identify an optimal edge weight and community detection algorithm. Several edge weights (unweighted, weighted by normalized number of projects, weighted by normalized value of projects) and community detection algorithms (Spinglass, Louvain, Fast Greedy) were considered. The optimal solution was chosen based on the modularity score and reasoning of the generated topic clusters.

This section describes how the collected data was used further in the next phase of development including network generation and topic and researcher clustering.

[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.

4.3.1 Networks of Topics

The concept of the topic-based network involved two potential ways that the network could be constructed. The topics could be analysed from the perspective of grants as well as researchers. This resulted in the creation of two different topic networks. In both networks, nodes represent topics. However, edges represent either grants or researchers. For ease of reference, the networks will be referred to as Grants as edges and Researchers as edges, respectively.

4.3.1.1 Grants as edges

This network consists of nodes representing topics and edges representing grants. An edge between two nodes means that two topics have a grant in common. Each grant record consists of a *EPSRC Research Topic Classifications* field consisting of research topics that classify the grant. In the creation process of this network, grants with only two or more topics were considered. For each grant record, an edge was "drawn" between each of the topics. Fig. 4.5 provides a visual explanation of the network's structure.

The network consists of two node and edge attributes, number and value of grants. The node and edge attributes are explained further in the *Node and Edge attributes* section.

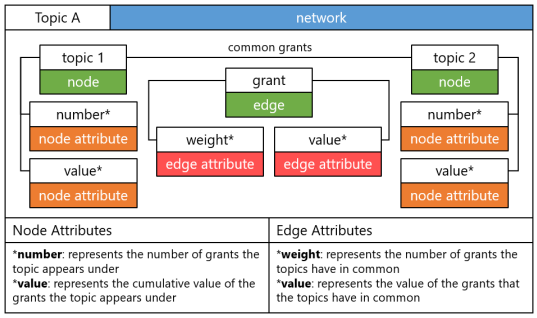


Figure 4.5: Visual explanation of the Topic network (Grants as edges).

4.3.1.2 Researchers as edges

This network consists of nodes representing topics and edges representing researchers. An edge between two nodes means that two topics have a re-

searcher in common. Each researcher record consists of a *Current EPSRC-Supported Research Topics* field consisting of the researcher's current research topics. In the creation process of this network, researchers with only two or more current topics were considered. For each researcher record, an edge was "drawn" between each of the topics. Fig. 4.6 provides a visual explanation of the network's structure.

The network consists of a node and edge attribute, number of researchers. The node and edge attributes are explained further in the *Node and Edge attributes* section.

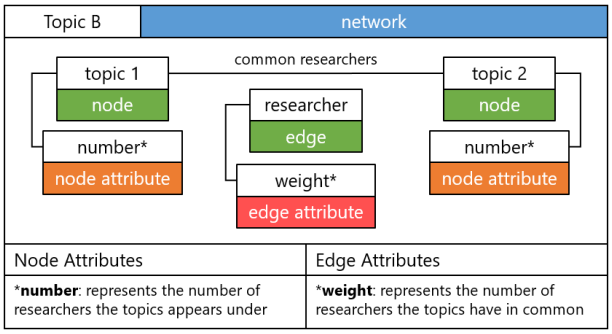


Figure 4.6: Visual explanation of the Topic network (Researchers as edges)

4.3.2 Networks of Researchers

The concept of the researchers network involved two potential ways that the network could be constructed. The researchers could be analysed from the perspective of grants as well as topics. This resulted in the creation of two different researcher networks. In both networks, nodes represent researchers. However, edges represent either grants or topics. For ease of reference, the networks will be referred to as Grants as edges and Topics as edges, respectively.

4.3.2.1 Grants as edges

This network consists of nodes representing researchers and edges representing grants. An edge between two nodes means that two researchers have a grant in common. Each grant record consists of a *EPSRC Research Topic Classifications* field consisting of research topics that classify the grant.

In the creation process of this network, grants with only two or more topics were considered. For each grant record, an edge was ”drawn” between each of the topics. Fig. 4.7 provides a visual explanation of the network’s structure.

The network consists of two node and edge attributes, number and value of grants. The node and edge attributes are explained further in the *Node and Edge attributes* section.

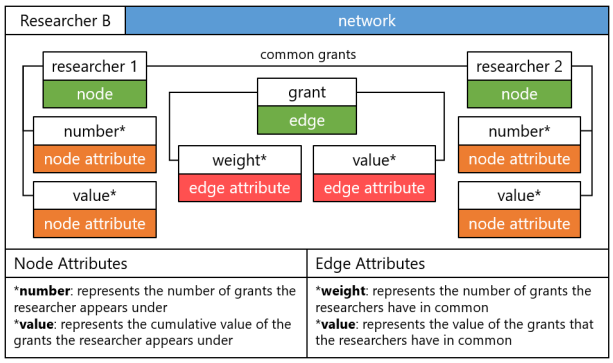


Figure 4.7: Structure of the Researcher network (Grants as edges)

4.3.2.2 Topics as edges

This network consists of nodes representing researchers and edges representing topics. An edge between two researchers means that two researchers have a topic in common. Each researcher record consists of a *Current EPSRC-Supported Research Topics* field consisting of the researcher’s current research topics. In the creation process of this network, each researcher was compared to all other researchers and if it was found that they had at least one topic in common, an edge was ”drawn” between the researchers. Fig. 4.8 provides a visual explanation of the network’s structure.

The network consists of a node and edge attribute, number of topics. The node and edge attributes are explained further in the *Node and Edge attributes* section.

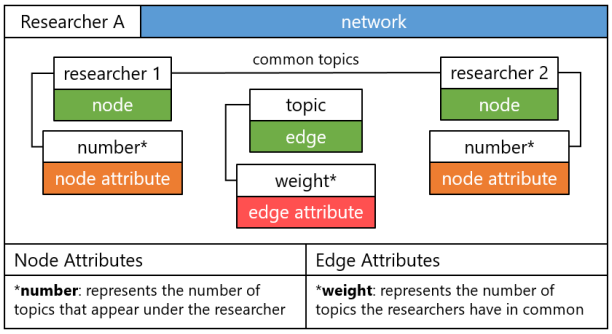


Figure 4.8: Structure of the Researcher network (Topics as edges)

4.3.3 The 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 Grants on the Web (GoW) service. First and foremost, the former is not unique within a network, while the latter is unique within the GoW service. This is because of the way links between topics or researchers were established in a network. For example, if a grant is classified by three topics, each topic is linked to one another by a separate edge, which represents the grant in question. Therefore, in this case, three edges are created to link the topics which represent the same grant, which means that the grant will appear within the network three times. It is also crucial to specify that unlike grants, a network consists of 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 an edge represents. It is only known that it represents one or more grants. If this was known, providing an answer to the above questions significantly easier. In contrast, the answer was achieved through a lengthier process. However, the edges linking topics within the same community and the edges linking topics in different communities were known. This greatly aided the identification process of both the number and value of grants within communities 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 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. A number and value counter was created to keep track of the number and value of grants. When the check was true, the reference and value of the grant were added as the key and value of 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 was also achieved.

4.3.4 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 and Researcher network (Grants as edges) consists of a number and value attribute, while the Topic (Researchers as edges) and Researcher (Topics as edges) 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. This allows the analysis and visualisation of the network from two different perspectives.

4.3.4.1 Number of grants/topics/researchers

The number attribute has a number of different contexts, depending on the network it's a part of. Firstly, in the Topic and Researcher networks (Grants as edges), the node number attribute represents the number of grants that contain a topic or researcher. In the same networks, the edge number attribute represents the number of grants two topics or researchers have in common, meaning they are both contained within the same grant. Secondly, in the

Topic network (Researchers as edges), the node number attribute represents the number of researchers that contain the topic, while the edge number attribute represents the number of researchers two topics have in common. Thirdly, in the Researcher network (Topics as edges), the node number attribute represents the number of topics a researcher has, while the edge number attribute represents the number of topics two researchers have in common.

4.3.4.2 Value of grants

Similarly, the value attribute also has a number of different contexts, depending on the network it's a part of. Firstly, networks that are not based on grant data do not contain of the value attribute. Secondly, in the Topic and Researcher network (Grants as Edges), the node value attribute represents the value of a topic or researcher. This means 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.

4.3.5 Normalization of node and edge attribute values

The numerical values used as the node and edge attributes, especially the value of grants attribute, represent very large numbers which caused some issues in terms of development, analysis and visualisation. To accomodate this, the values underwent a normalization process which scaled the value range down. The formula used to normalize the values is presented below:

$$(val - old\_min) \times new\_range / old\_range) + new\_min) \tag{4.1}$$

where:

- *val* is the value being normalized
- *old_min* is the minimum of the initial value range
- *new_range* is the new range values will be normalized to

- *old_max* is the maximum of the initial value range
- *new_min* is the minimum of the new value range

4.4 Testing of edge weights and algorithms

In the previous section, the process carried out to normalize the values used as node and edge attributes was described. This section describes the experiments carried in order to identify the most reasonable division of the networks of topics and researchers into clusters of topics and researchers.

Several experiments were carried seeking to identify an optimal edge weight and community detection algorithm that produced the most rational clustering results. In order to ensure a consistent analysis, the experiments were solely carried out on the Topic network (Grants as edges). The results of the experiments on this network were applied to all the other networks.

Eight different community detection algorithms were considered including Louvain, Spinglass and Fast Greedy. The Topic network (Grants as edges) consists of two node and edge attributes, number and value of grants.The values of both attributes are significantly large and therefore they were normalized for development, analysis and visualisation purposes. In total, five different "candidates" of interpreting the edge weight were considered: unweighted, weighted by normalized number of grants and weighted by normalized value of grants.

Furthermore, eight community detection algorithms were applied to the network using each of the three edge weight interpretations. Subsequently, the results of the experiments underwent a multi-phase comparative analysis which is documented in the next section.

4.5 Identification of an optimal edge weight and community detection algorithm

The experimental stage consisted of a number phases. In the first phase, a number of community detection algorithms were applied to the network

using each of the edge weight interpretations. The resulting modularity scores and 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 interpretation was eliminated in the first phase. Furthermore, a number of community detection algorithms such as Label Propagation, Leading Eigenvector and Edge Betweenness were also dropped due to low results. In contrast, the Louvain and Spinglass algorithms remained to be considered further.

During the second phase, the number of topics within each community was compared across all candidates, and it was decided that all remaining edge weight interpretation and community detection algorithms should proceed to the final testing phase. In the third phase, the final phase, each community of topics identified using each one of the edge weight interpretations and community detection algorithms were compared to each other with rationality as a criteria. Finally, it was concluded that the clustering produced by the normalized number of grants edge weight interpretation and the Louvain community detection algorithm was the most reasonable and the one that would be used for the rest of the analysis.

4.6 Clustering of Topics and Researchers

As revealed in the previous section, the Louvain community detection algorithm performed the best when applied to a network in which the weight of an edge represents the normalized number of grants, topics or researchers. Following this discovery, the combination of edge weight and community detection algorithm was considered an optimal solution.

Furthermore, the solution was applied to each network in order to identify its underlying community structure. This resulted in a number of communities also known as clusters or groups, consisting of topics or researchers. Each topic or researcher is assigned to exactly one cluster. Cer-

tain identified communities can be quite broad in terms of representing a few research areas, especially if the network consists of a large number of nodes. In order to achieve a more specific perspective, the community detection algorithm was applied to each community in order to identify its sub-communities which form its underlying community structure.

Moreover, the topic and researcher clustering was documented in a spreadsheet using Microsoft Excel. The number and value of grant records in each community was also documented and calculated. This was followed by the creation of word cloud representations of the communities and sub-communities identified within the Topic network (Grants as edges), using Wordle. Each community and sub-community was illustrated from three different perspectives: word frequency, number of grants and value of grants.

4.7 Evaluation of Topic and Researcher clusters

Once the topic and researcher clusters were identified and defined, they went through and evaluation stage which determined how strong the links between topics in the same clusters are compared to links that connect topics that are in different clusters. This was achieved by calculating the Dice and Jaccard similarity of topics within the same and different clusters.

Firstly, the origin and destination nodes of each edge in the network were identified. This forms a pair of nodes: origin node, destination node. Certain edges connect topics that are in the same cluster, while others link topics that are in different clusters. Therefore, some node pairs represent topics from the same cluster, while others represent topics from different clusters. Secondly, the Dice and Jaccard similarity between each node pair. Theoretically, the similarity between a pair of nodes from the same cluster should be higher than the similarity between a pair of nodes from different clusters.

Finally, in order to get an overall perspective, the average Dice and Jac-

card similarity of nodes within and between clusters was calculated. The evaluation phase helped to determine whether in reality this is also true.

4.8 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.

4.8.1 JetBrains PyCharm used for programming

JetBrains PyCharm [16] is an Integrated Development Environment (IDE) for programming in Python. It also provides support for writing code in Bash. This tool was used to write the development code in primarily Python but also in Bash.

4.8.2 Microsoft Excel used for data storage

Microsoft Excel [17] is a spreadsheet software featuring calculation, graphing tools, pivot tables, and macro programming support in Visual Basic. This tool was used to explore, filter and validate the collected data.

4.8.3 iGraph used for network analysis and visualisation

iGraph [18] is a library collection for creating and manipulating graphs and analyzing networks, written in C and also available as packages for Python and R. This tool was used to compute the majority of the network properties as well as applying several community detection algorithms. It was also used to produce plots of the networks created.

4.8.4 NetworkX used for visualising adjacency matrices

NetworkX [19] is a package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks in Python. This tool was used to visualise adjacency matrices.

4.8.5 Wordle used for creating word cloud visualisations

Wordle [20] 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 control the font size as well as specifying a colour for each word. This project used Wordle to create word clouds representing the topic clusters using the number and value attributes to control font size, and the different clusters to control colour.

4.8.6 Adobe Photoshop used for image editing

Adobe Photoshop [21] is a graphics editor developed by Adobe Systems. This tool was used for image editing as well as turning network plots produced by iGraph into complete network visualisations.

5 Results and Evaluation

The previous section described the methods and tools used throughout a number of processes including data collection, network creation and testing. It documents the journey of collecting data and interpreting it in a way that allowed the creation of a number of networks.

This study is very distinctive in terms of the data used. Other network science projects will make use of data that already carries a suitable network format, which enables its analysis straight away. However, in this case, as the approach to the problem is a novel one, the data has not been pre-processed for network analysis. Therefore, part of the project and its objective concentrates on data collection and formatting in order to be able to create a number of network which can then be analysed. Furthermore, this means that the networks constructed from the collected EPSRC data are part of the results of this project.

This section presents the results produced by the project while also covering their evaluation. The results including the networks constructed from the collected data as well as the clusters of topics and researchers identified. The current clusters are compared with historical cluster which were also identified in order to discover transitional trends related to research and funding. Furthermore, different network interpretations of the data are also compared in order to identify which one yields a more rational clustering.

5.1 Networks of Topics

One of the subsections of the Methodology section describes the methods used in order to create a number of different networks from the collected data. This section presents and describes the results of following those methods.

One of the types of network created is a Network of Topics. It consists of nodes which represent topics and edges which have two interpretations, as they can either represent grants or researchers. A separate network was con-

structed to fit each interpretation. They are referred to as Grants as edges and Researchers as edges. Moreover, as historical grant data was also collected, a further two versions of the Grants as edges Topic network were constructed, which leads to a total of four networks constructed.

5.1.1 Grants as edges

The Grants as edges Topic network constructed from current grant data consists of 223 nodes representing as many topics and 2008 edges representing common grants between topics. In comparison, both historical networks have less nodes with the 2010-2000 and 2000-1990 networks consisting of 208 and 136 nodes, respectively. This is expected, as the number of disciplines was lower, gradually increasing with time. Interestingly, the 2010-2000 network consists of more edges which transitions into more topics being connected to each through common grants. However, the increased number of edges also seems to have some correlation to the fact that the network is unconnected, while the other two are connected. This means that an edge does not exist between every pair of nodes. Furthermore, all networks are weighted and the edge weight represents the number of grants that two topics have in common. Table 5.1 presents the full set statistics of the Topic network (Grants as edges).

Table 5.1: Statistics of the Topic network (Grants as edges)

	Current	**2010-2000**	**2000-1990**
Nodes	223	208	136
Edges	2008	3592	748
Type	Undirected	Undirected	Undirected
Weighted	Yes	Yes	Yes
Average Degree	18.009	34.538	11
Average Weighted Degree	19.543	35.337	12.721
Diameter	5	5	6
Density	0.081	0.167	0.081
Modularity	0.373	0.271	0.4
Weak Components	1	2	1
Average Clustering Coefficient	0.597	0.59	0.453
Average Path Length	2.395	2.077	2.54

5.1.2 Researchers as edges

The Researchers as edges Topic network represents the second interpretation of the data, and is formed of topics represented by 225 nodes and 5192 edges. The substantial increase in edges compared to the Grants as edges interpretation, is in part due to the fact that the number of researcher records exceeds the number of grant records by 2699 records. This network is also weighted, but this time, the edge weight represents the number of researchers two topics have in common. Table 5.2 presents the full set of statistics of the Topic network (Researchers as edges).

Table 5.2: Statistics of the Topic network (Researchers as edges)

	Current
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
Weak Components	1
Average Clustering Coefficient	0.715
Average Path Length	1.925

5.2 Clusters of Topics

The previous section presented part of the first batch of results that the project produced which is the Network of Topics. The optimal solution identified at the end of the experimental phase was applied to the networks constructed, which resulted in a number topic clusters. This section presents and details the results for Topic networks, Grants and Researchers as edges.

5.2.1 Grants as edges

The application of the Louvain community detection algorithm on the Topic network (Grants as edges) resulted in the identification of 6 different com-

munities of topics. Table 5.3 presents the number of nodes representing topics, the number and value of grants and the predominant words within each community. The predominant words were identified based on the word frequency using Wordle. The complete clustering of the Topic network (Grants as edges) is presented in Table part of Appendix B.

Table 5.3: Number of nodes and grants, value of grants and the predominant words based on word frequency of each community identify within the current Topic network (Grants on edges) 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 network (Grants on edges).

	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 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 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 £192 million. This is justified as the topic also appears in 155 grants, more than any other in Community 1. Figure 5.1 presents a word cloud representation of Community 1.

Community 2 is not as well defined as Community 1 as it consists of several topics which have obvious differences including *product design*, *artificial intelligence*, *developmental psychology*, *human communication in ict*, *criminal law & criminology*, and *comput./corpus linguistics*. This is justified considering the

Figure 5.1: Topics clustered within Community 1. Font size represents the number of grants that contain the topics.

significant size of the community. However, there are grants which represent a study involving a combination of two topics which are somewhat different such as *artificial intelligence* and *linguistics*, for example. *Natural language processing* is a field of both *artificial intelligence* and *computational linguistics*. Figure 5.2 presents a word cloud representation of Community 2.

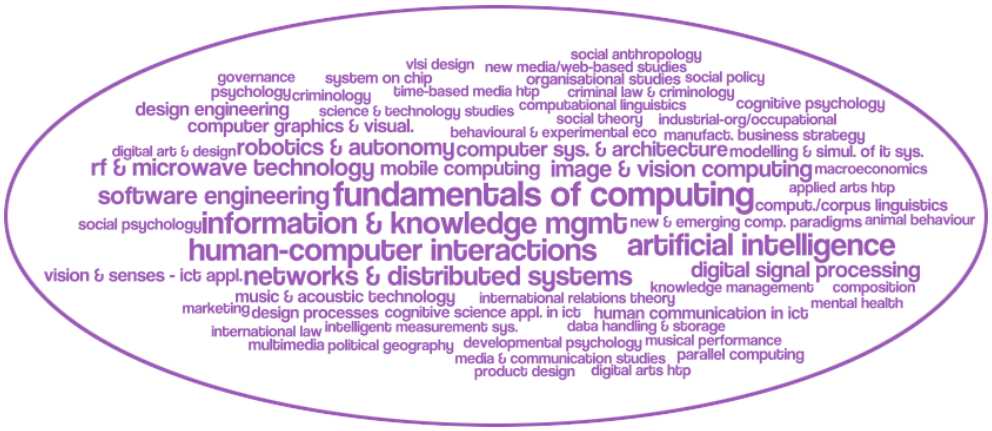


Figure 5.2: Topics clustered within Community 2. Font size represents the number of grants that contain the topics.

Community 3 also represents a comprehensible clustering enclosing topics such as *fluid dynamics*, *microsystems* and *wind power*. It consists of three major topics both in term of number of grants that contain them and the value of the grants: *materials characterisation* (270 grants, worth £350M), *materials synthesis & growth* (273 grants, worth £305M) and *manufacturing machine & plant* (196 grants worth £273M). Figure 5.3 presents a word cloud representation of Community 3.

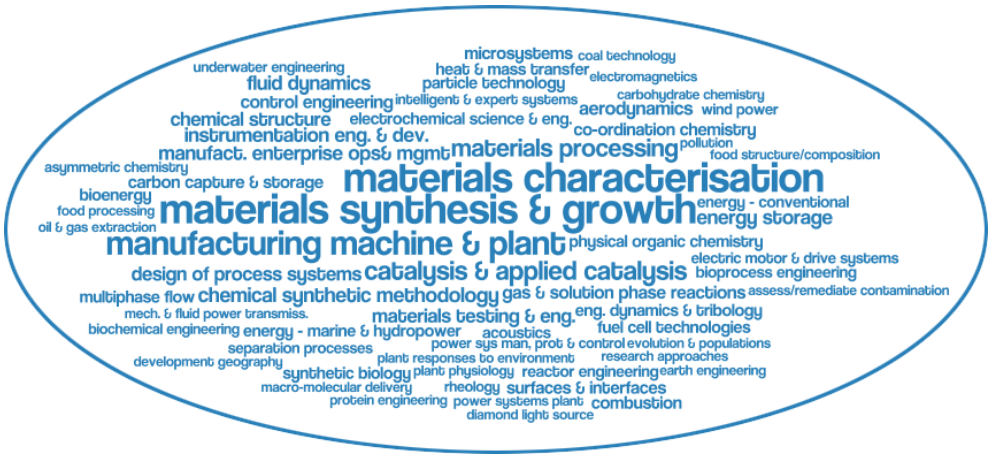


Figure 5.3: Topics clustered within Community 3. Font size represents the number of grants that contain the topics.

Community 4 is the smallest community in size (10 topics) but also the most well-defined community, as the topics clustered within it have a 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 were involved in most grants, 131 and 120, respectively. In terms of value, the latter is valued higher than the former, £200M compared to £69M. Figure 5.4 presents a word cloud representation of Community 4.

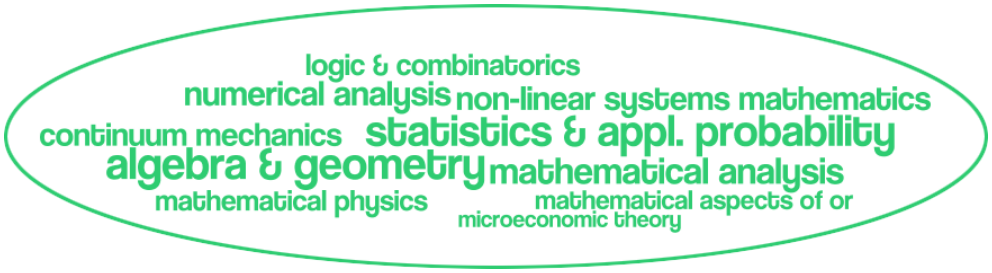


Figure 5.4: Topics clustered within Community 4. Font size represents the number of grants that contain the topics.

Community 5 represents a coherent clustering of topics including *energy efficiency*, *geohazards*, *environment & health* and *urban & land management*. This community represents topics from different fields which are contained in grants which aim to tackle a common environment-related problem such as climate change or global warming. The most popular topic in terms of num-

ber of grants is *energy efficiency*, with 100 grants valued at £150M. Figure 5.5 presents a word cloud representation of Community 5.

Figure 5.5: Topics clustered within Community 5. Font size represents the number of grants that contain the topics.

Community 6 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 involved in the highest number of grants, 82 worth £97M. Figure 5.6 presents a word cloud representation of Community 6.

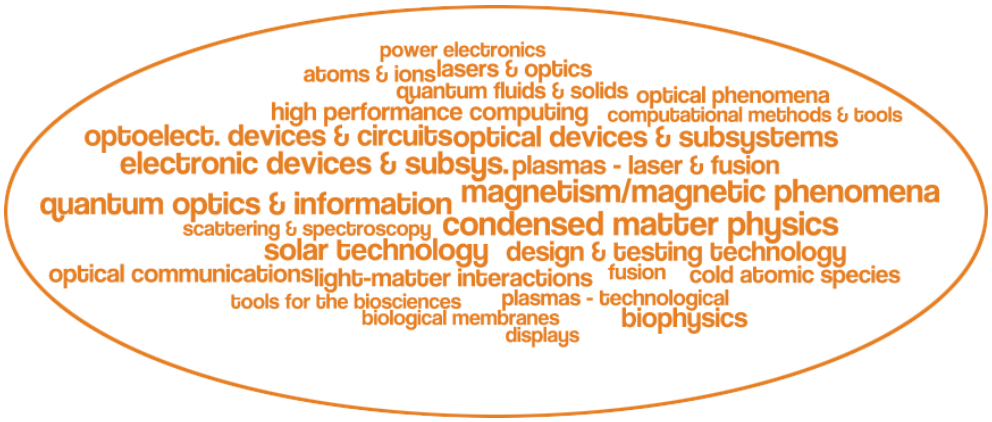


Figure 5.6: Topics clustered within Community 6. Font size represents the number of grants that contain the topics.

5.2.1.1 Historical data comparison

Over the years, the research trend led to new topics being defined, while others were discontinued as they were not necessary anymore. For example, *bionanoscience*, *escience* and *language acquisition* are topics which existed between 2000 and 2010 but do not exist currently. In contrast, *animal organisms*, *criminal law & criminology* and *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 3072 grants within communities with a total value of £3.5B. Between, 2010-2000, researchers worked on 16617 grants, valued at £4.9B. These figures indicate a significant different 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 received significantly more funding.

Furthermore, this is supported by the number and value of the grants from 1990 to 2000. Researchers worked on a slightly less number of grants than between 2000 and 2010, but they also received significantly less funding, £1.7B.

In terms of clustering, the historical communities identified within the Topic network (Grants as edges) hold slight differences when compared to the current communities. First and foremost, the community detection algorithm identified 5 communities in both historical networks. This can be the results of the contrast in the number of topics and the actual topics between the current and historical networks. Moreover, the most well-defined, Mathematics-focused community (Community 4) identified in the current network, is not well-defined anymore, as it is part of a larger community also including engineering (Community 2). This symbolises the current increase in the number of grants that are focused on *mathematics* only rather

than a combined effort with another topics such as *engineering*.

Furthermore, between 2000 and 2010, there was only one grant classified by *soil science* and *crop science*. In the current list of topics, *crop science* is not present anymore. Perhaps, its removal could be justified by the similarity between the topics, deeming one of them as unnecessary. This may also explain why the two topics didn't form a community in the current network.

Table 5.4: Number of nodes and grants, value of grants and the predominant words based on word frequency of each community identify within the historical (2000-2010) Topic network (Grants on edges). 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 (2000-2010) Topic network (Grants on edges).

	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

5.2.2 Researchers as edges

The application of the Louvain community detection algorithm on the current Topics network (Researchers as edges) resulted in the identification of 4 different communities of topics, 2 communities less than within the Topic network (Grants as edges). Note that the Topic network (Researchers as edges) was constructed using the current data set only, because the researcher records within EPSRC only provide a researcher's current topics, and not past topics. Table 5.6 presents the number of nodes representing researchers as well as the predominant words within each community. The predominant words were identified based on word frequency using Wordle. The complete clustering of the current Topic network (Researchers as edges)

Table 5.5: Number of nodes and grants, value of grants and the predominant words based on word frequency of each community identify within the historical (1990-2000) Topic network (Grants on edges) 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 valueof unique grants in communities within the historical (1990-2000) Topic network (Grants on edges).

	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

is presented in Table part of Appendix B.

Table 5.6: Number of nodes and grants, value of grants and the predominant words based on word frequency of each community identify within the current Topic network (Researchers as edges) The number of grants includes duplicate grants, as a grant can be contained by 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 network (Researchers as edges).

	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

5.2.3 Comparison of Grants and Researchers as edges

There are obvious differences between the clustering produced using the Topic network (Grants as edges) and Topic network (Researchers as edges). Firstly, the number of communities identified differs, as using the former

the Louvain community detection algorithm identified 6 communities, compared to 4 identified using the latter. This results in an imbalance in community size, with one of the communities (Communities identified within the Topic network (Researcher as edges) network consisting of 109 topics. A large community is also a broad community, which means that is less specific and lacks a clear representation of a research areas. It also causes other communities to be significantly smaller in size.

Furthermore, using the this network, the Mathematics-based community is larger in size and not as well-defined including irrational topic additions such as *genomics*. Moreover, the clustering produced using this network also lacks the identification of the Biology-focused community. However, there are also similarities between the communities identified using the two types of networks, as the *engineering* and *chemistry*-based communities appear in both.

In conclusion, this comparative analysis showed that the clustering produced using the Topic network (Grants as edges) is more detailed and rational compared to the one that produced using the Topic network (Researchers as edges) network.

5.2.4 Evaluation of Topic clusters

5.2.4.1 Grants as Edges

This section showcases the results of the evaluation phase. Pairs of nodes that are both from the same cluster and different clusters were identified. Moreover, the Average dice and Jaccard similarity between both types of node pairs was calculated. Table 5.7 presents the results of the evaluation phase carried out on the current and historical Topic networks (Grants as edges). The results show that nodes within the same cluster have a higher similarities than nodes from different clusters. This outcome also translates to the historical networks.

Table 5.7: Dice and Jaccard similarity coefficients of node pairs within and between communities in the current and historical Topic networks (Grants as edges). Each node pair represents an edge which connects two nodes within the same community or in two different communities. If the network division is strong, it is expected that a node pair within the same community should be more similar compared to a node pair consisting of nodes in two different communities. IN stands for within communities, while OUT means between communities.

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

5.2.4.2 Researchers as Edges

This section showcases the results of the evaluation phase. Pairs of nodes that are both from the same cluster and different clusters were identified. Moreover, the Average dice and Jaccard similarity between both types of node pairs was calculated. Table 5.8 presents the results of the evaluation phase carried out on the current Topic networks (Researchers as edges). The results show that nodes within the same cluster have a higher similarities than nodes from different clusters.

5.3 Networks of Researchers

5.3.1 Grants as edges

The Grants as edges Researcher network is a carbon copy of the Grants as edges Topic network, with the exception that nodes represent Researchers and not Topics. Its current data-based form is composed of 260 nodes and 208 edges representing one or more common grants between two researchers. Furthermore, its historical data-based form consists of both an increased number of nodes and edges. The network is weighted, with its

Table 5.8: Dice and Jaccard similarity coefficients of node pairs within and between communities in the current Topic network (Researchers as edges). Each node pair represents an edge which connects two nodes that may be within the same community or in two different communities. If the network division is strong, it is expected that a node pair within the same community should be more similar compared to a node pair consisting of nodes in two different communities. IN stands for within communities, while OUT means between communities.

	Current
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

edge weight attribute representing the number of grants that two researchers have in common. Table 5.9 presents the full set of statistics of the Researcher network (Grants as edges).

Table 5.9: Statistics of Researcher network (Grants as edges)

	Current	**2010-2000**	**2000-1990**
Nodes	260	2434	1847
Edges	208	4919	2002
Type	Undirected	Undirected	Undirected
Weighted	Yes	Yes	Yes
Average Degree	1.60	4.042	2.168
Average Weighted Degree	2.592	7.794	2.798
Diameter	16.0	36.0	18.0
Density	0.006	0.002	0.001
Modularity	0.955	0.977	0.978
Weak Components	89	473	559
Average Clustering Coefficient	0.578	0.748	0.648
Average Path Length	1.942	6.874	3.676

5.3.2 Topics as edges

The Topics as edges Researcher network is a reversed version of the Researchers as edges Topic network. It consists of 655 researchers linked by 4548 edges representing common topics between the researchers. The network is undirected and weighted with the edge weight representing the number of topics two researchers have in common. In comparison to current version of the Grants as edges Researcher network which consists of 260 nodes and 208 edges, this network consists of substantially more nodes and edges. Table 5.10 presents the full set of statistics of the Researcher network (Topics as edges).

Table 5.10: Full set of statistics of the Researcher network (Topics as edges)

	Current
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
Weak Components	39
Average Clustering Coefficient	0.825
Average Path Length	4.278

5.4 Clusters of Researchers

The previous two sections presented part of the first and second batch of results that the project produced which is the Networks of Topics and the Clusters of Topics. The optimal solution identified at the end of the experimental phase was applied to the networks constructed, which also resulted in a number researcher clusters. This section presents and details the results for both Researcher networks, Grants and Topics as edges.

5.4.1 Grants as edges

The application of the Louvain community detection algorithm on the Researcher network (Grants as edges) resulted in the initial and excessive identification of 89 communities of researchers. This is due to the lack of strong and frequent relationships between different researchers. Note that the Researcher network (Grants as edges) consists of researchers linked by their common grants. Also note that, during the network creation process, due to the magnitude of the network, the Researcher network (Grants as edges) was sampled and only includes researchers that have at least two grants in common. This represents a fairly strong collaboration link between two researchers.

In conclusion, it seems that two or more researchers collaborating multiple times as part of a grant is a rarity within the EPSRC data. Due to the sparse nature of the communities identified, the two largest communities are presented here. Table 5.3 presents the number of nodes representing researchers within each community. The complete clustering of the Researcher network (Grants as edges) is presented in Table part of Appendix B.

Table 5.11: Number of researchers and grants and value of grants in each community identified within the current Researcher network (Grants on edges) 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 *Total* column represents the number and value of unique grants in communities within the current Researcher network (Grants on edges).

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

5.4.1.1 Historical data comparison

Similar to the historical data comparison of the Topic network, there is a clear funding trend which increases more rapidly the more recent the grant is. Currently, there are 65 grants within the two largest communities worth

a total value of Âč3.5B.

Between, 2010-2000, researchers worked on 866 grants, valued at £551M. 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 received significantly more funding. Furthermore, this is supported by the number and value of the grants from 1990 to 2000. Researchers worked on a slightly less number of grants than between 2000 and 2010, but they also received significantly less funding, £130M.

Table 5.12: Number of researchers and grants and the value of grants of each community identified within the historical (2000-2010) Researcher network (Grants on edges). 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 *Total* column represents the number and value of unique grants in communities within the historical (2000-2010) Researcher network (Grants on edges).

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

Table 5.13: Number of nodes and grants, value of grants and the predominant words based on word frequency of each community identify within the historical (1990-2000) Researcher network (Grants on edges) 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 *Total* column represents the number and value of unique grants in communities within the historical (1990-2000) Researcher network (Grants on edges).

	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

5.4.2 Topics as edges

In contrast to the Researcher network (Grants as edges), the application of the Louvain community detection algorithm on the Researcher network (Topics as edges) 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.

Note that the Researcher network (Topics as edges) consists of researchers linked by their common topics. Also note that, during the network creation process, due to the magnitude of the network, the Researcher network (Topics as edges) was sampled and only includes researchers that have at least five topics in common. This represents a strong topic sharing link between two researchers.

In conclusion, 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 5.3 presents the number of nodes representing researchers within each community. The complete clustering of the Researcher network (Topics as edges) is presented in Table part of Appendix B.

Table 5.14: Number of researchers of each community identify within the current Topic network (Researchers as edges) The number of grants includes duplicate grants, as a grant can be contained by more than one community. Subsequently, the value of grants also includes the duplicate grants. However, the *Total* column represents the number and value of unique grants in communities within the current Researcher network (Topics as edges).

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

5.4.3 Comparison of Grants and Topics as edges

The results produced using both networks clearly indicate that the Researcher network (Topics as edges) is the more rational and successful in-

terpretation of the data. Prior to the network creation process, this was expected.

In one hand, 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. It is essential to note that there is a clear difference between collaborating on a government-funded grant and a research paper. 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 funding over a 26-year long period.

One the other hand, a network that consists of researchers and links between them symbolising a shared research interest the two. This type of connection between two researchers is not limited to sharing a relationship, institution or organisation or location. The only thing needed is a a shared research interest. In the academic world, this is very common as researchers from all over the world can related to each other through a common research topic or 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 is also large and at the same time sparse. That being said, this interpretation of the data is still an approach involving more reasoning on the available research data.

5.4.4 Evaluation of Researcher clusters

5.4.4.1 Grants as Edges

This section showcases the results of the evaluation phase. Pairs of nodes that are both from the same cluster and different clusters were identified.

Moreover, the Average dice and Jaccard similarity between both types of node pairs was calculated. Table 5.15 presents the results of the evaluation phase carried out on the current and historical Researcher networks (Grants as edges). The results show that nodes within the same cluster have a higher similarities than nodes from different clusters. This outcome also translates to the historical networks.

Table 5.15: Dice and Jaccard similarity coefficients of node pairs within and between communities in the current and historical Researcher networks (Grants as edges). Each node pair represents an edge which connects two nodes that may be within the same community or in two different communities. If the network division is strong, it is expected that a node pair within the same community should be more similar compared to a node pair consisting of nodes in two different communities. IN stands for within communities, while OUT means between communities.

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

5.4.4.2 Topics as Edges

This section showcases the results of the evaluation phase. Pairs of nodes that are both from the same cluster and different clusters were identified. Moreover, the Average dice and Jaccard similarity between both types of node pairs was calculated. Table 5.16 presents the results of the evaluation phase carried out on the current Researcher network (Topics as edges). The results show that nodes within the same cluster have a higher similarities than nodes from different clusters.

Table 5.16: Dice and Jaccard similarity coefficients of node pairs within and between communities in the current Researcher network (Topics as edges). Each node pair represents an edge which connects two nodes that may be within the same community or in two different communities. If the network division is strong, it is expected that a node pair within the same community should be more similar compared to a node pair consisting of nodes in two different communities. IN stands for within communities, while OUT means between communities.

	Current
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

6 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 surrounding EPSRC, the concept of modularity and community detection algorithms was provided. Works that contributed to the project were also described and their contribution outlined. Additionally, the methods put into practice throughout every stage of the project. were explained extensively.

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

The experimental comparison phase resulted in a significant and valuable set of results. Firstly, edges weighted by the normalized number of grants was determined as the optimal interpretation of the edge weight attribute due to its high modularity score and reasonable clustering produced.

Secondly, the Louvain community detection algorithm developed by Blondel et al., was "crowned" as the optimal community detection method due to its high performance in the experiments and the well-defined nature of the community structure it identified. Finally, the Topic (Grants as edges) and Researcher (Topics as edges) networks proved to be better solutions compared to the other network interpretations as they produced more balanced and rational clusters of topics and researchers.

This thesis, based to the knowledge available, represents the first approach of deploying graph theory in order to provide an solution to a real-world problem which at the moment, is specific to one organisation. With the data undergoing extensive experimental comparison, an evaluated solution boasting surprisingly high performance is identified and provided while its benefits and limitations are also outline. Due to the short amount of time available, the analysis performed on the researcher data is slightly limited in terms of the quantity of data collected and the load of work carried on it. However, this is justified by the lower value that the researcher network brings in comparison to the topic data. This is discussed further in the the next section.

In conclusion, this project represents clear evidence that the novel approach based on graph theory is of great value and because due to the fact that it isn't limited by any data set, it could be used to seek solutions to other real-world problems such as the identification of topic communities within a network of newspaper articles.

6.1 Further potential 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 were identified:

- Extending comparison experiments phase to include more community detection algorithms

- Extending the analysis of Networks of Researchers
- 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 phase which is 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 tested. 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 aren't part of iGraph. Therefore, these algorithms could be implemented and added to the tests in order to determine whether better results than the ones produced by the Louvain community detection algorithm can be achieved.

Secondly, the analysis of the Networks of Researchers 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 while researchers aren't unless those same people know the researchers personally or work within EPSRC.

However, the EPSRC Grants on the Web (GoW) service provides substantial public data, which is partly used in this project, but not fully. Both researcher and grants records consist of additional information which is not used represented by fields such as the department, organisation, 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 and will increase the level of contextual information available and provide a guaranteed extension of analysis scope.

Finally, another way of extending this work can involve a comparative analysis of the results produced throughout this project and the results of a study with a similar objective 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. A study revolving around such an analysis would provide valuable insights into whether there is a potential correlation between the two sets of results while it will also help cement the use of graph theory as a optimal and valuable solution to solving clustering problems within both real-world and scientific scenarios.

A Data collected from EPSRC

A.1 Networks of Topics

A.1.1 Grants as edges

A.1.1.1 Current data set (2010 to 2016)

Table A.1: Topics and the number and value of grants that contain each topic sorted by the number of grants (largest to smallest) in the Topic network (Grants as edges) constructed using the current data set (2010 to 2016)

Topic	Number of Grants	Value of Grants
materials synthesis & growth	273	£305,344,034.00
materials characterisation	270	£350,845,612.00
manufacturing machine & plant	196	£273,265,770.00
fundamentals of computing	175	£140,728,848.00
med.instrument.device& equip.	155	£192,497,214.00
human-computer interactions	147	£193,369,943.00
artificial intelligence	146	£231,826,208.00
information & knowledge mgmt	141	£211,234,642.00
algebra & geometry	131	£69,459,955.00
catalysis & applied catalysis	124	£116,914,672.00
statistics & appl. probability	120	£200,779,552.00
networks & distributed systems	110	£177,563,885.00
materials processing	108	£203,775,655.00
energy efficiency	100	£150,990,954.00
analytical science	90	£145,977,721.00
software engineering	87	£92,094,590.00
mathematical analysis	86	£56,207,447.00
condensed matter physics	82	£97,207,707.00
biomaterials	81	£101,030,647.00
rf & microwave technology	81	£71,900,374.00
energy - nuclear	79	£77,162,994.00

Table A.1: Topics and the number and value of grants that contain each topic sorted by the number of grants (largest to smallest) in the Topic network (Grants as edges) constructed using the current data set (2010 to 2016)

Topic	Number of Grants	Value of Grants
robotics & autonomy	77	£93,009,453.00
image & vision computing	74	£116,964,229.00
magnetism/magnetic phenomena	74	£61,078,943.00
chemical synthetic methodology	72	£101,788,856.00
electronic devices & subsys.	72	£123,197,049.00
numerical analysis	68	£68,652,593.00
quantum optics & information	68	£99,752,497.00
energy storage	65	£71,775,478.00
instrumentation eng. & dev.	65	£68,148,151.00
manufact. enterprise ops& mgmt	65	£119,063,980.00
non-linear systems mathematics	62	£75,268,389.00
sustainable energy networks	62	£91,222,489.00
digital signal processing	61	£74,370,433.00
fluid dynamics	61	£51,953,657.00
solar technology	61	£40,262,413.00
medical imaging	60	£73,682,425.00
continuum mechanics	57	£56,547,841.00
materials testing & eng.	57	£94,707,839.00
complex fluids & soft solids	54	£51,883,688.00
optoelect. devices & circuits	54	£117,816,421.00
optical devices & subsystems	53	£99,280,018.00
chemical structure	52	£57,198,101.00
computer sys. & architecture	52	£60,302,509.00
design of process systems	52	£97,558,106.00
aerodynamics	50	£51,110,961.00
biomechanics & rehabilitation	47	£62,830,082.00
mobile computing	46	£67,914,107.00
tissue engineering	46	£85,152,510.00

Table A.1: Topics and the number and value of grants that contain each topic sorted by the number of grants (largest to smallest) in the Topic network (Grants as edges) constructed using the current data set (2010 to 2016)

Topic	Number of Grants	Value of Grants
biophysics	45	£44,234,268.00
control engineering	43	£66,051,915.00
logic & combinatorics	43	£24,297,595.00
design & testing technology	40	£63,777,809.00
gas & solution phase reactions	40	£47,267,022.00
urban & land management	39	£73,439,680.00
chemical biology	38	£64,182,393.00
co-ordination chemistry	38	£25,587,639.00
complexity science	38	£77,295,542.00
mathematical physics	38	£27,981,443.00
plasmas - laser & fusion	38	£32,160,232.00
transport ops & management	37	£60,741,338.00
computer graphics & visual.	36	£68,214,771.00
design engineering	36	£54,318,218.00
microsystems	36	£57,061,520.00
surfaces & interfaces	36	£33,940,347.00
structural engineering	35	£38,256,069.00
water engineering	35	£64,215,111.00
carbon capture & storage	34	£36,659,546.00
drug formulation & delivery	34	£55,765,735.00
building ops & management	33	£67,993,377.00
combustion	33	£27,342,322.00
ground engineering	33	£32,605,304.00
civil engineering materials	31	£19,613,822.00
coastal & waterway engineering	31	£22,329,706.00
electrochemical science & eng.	31	£30,783,229.00
sustainable energy vectors	31	£63,410,664.00
eng. dynamics & tribology	30	£36,426,390.00

Table A.1: Topics and the number and value of grants that contain each topic sorted by the number of grants (largest to smallest) in the Topic network (Grants as edges) constructed using the current data set (2010 to 2016)

Topic	Number of Grants	Value of Grants
medical science & disease	30	£56,249,992.00
synthetic biology	30	£52,068,052.00
bioenergy	29	£31,787,943.00
biological & medicinal chem.	29	£45,503,257.00
vision & senses - ict appl.	29	£25,991,424.00
fuel cell technologies	28	£41,944,010.00
energy - marine & hydropower	27	£34,863,003.00
heat & mass transfer	27	£37,435,379.00
high performance computing	27	£8,615,102.00
physical organic chemistry	27	£20,341,238.00
particle technology	26	£32,455,830.00
light-matter interactions	25	£35,116,928.00
multiphase flow	25	£24,773,814.00
optical communications	25	£60,200,016.00
lasers & optics	24	£59,840,920.00
mathematical aspects of or	24	£36,307,556.00
acoustics	23	£15,766,542.00
cold atomic species	23	£30,906,872.00
human communication in ict	23	£26,806,703.00
reactor engineering	21	£20,306,317.00
energy - conventional	20	£37,659,404.00
bioprocess engineering	19	£29,798,746.00
music & acoustic technology	19	£28,310,628.00
biomedical neuroscience	18	£29,394,374.00
separation processes	18	£18,158,126.00
wind power	17	£27,401,527.00
bioinformatics	16	£52,029,825.00
comput./corpus linguistics	14	£6,931,002.00

Table A.1: Topics and the number and value of grants that contain each topic sorted by the number of grants (largest to smallest) in the Topic network (Grants as edges) constructed using the current data set (2010 to 2016)

Topic	Number of Grants	Value of Grants
design processes	14	£22,405,461.00
atoms & ions	13	£6,214,215.00
construction ops & management	13	£24,891,070.00
electric motor & drive systems	13	£11,263,500.00
modelling & simul. of it sys.	13	£21,688,225.00
optical phenomena	13	£24,315,772.00
quantum fluids & solids	12	£12,041,645.00
asymmetric chemistry	11	£13,640,958.00
parallel computing	11	£12,635,358.00
rheology	11	£4,582,687.00
manufact. business strategy	10	£11,140,214.00
fusion	9	£190,002,289.00
management & business studies	8	£17,784,633.00
multimedia	8	£31,132,743.00
plasmas - technological	8	£10,160,650.00
psychology	8	£9,857,802.00
theoretical biology	8	£32,975,270.00
cognitive psychology	7	£3,369,386.00
cognitive science appl. in ict	7	£7,657,173.00
new & emerging comp. paradigms	7	£17,376,463.00
biochemical engineering	6	£18,986,493.00
social psychology	6	£4,160,146.00
soil science	6	£2,458,999.00
system on chip	6	£12,898,174.00
underwater engineering	6	£7,492,407.00
waste management	6	£12,104,515.00
waste minimisation	6	£8,717,814.00
bioelectronic devices	5	£22,042,473.00

Table A.1: Topics and the number and value of grants that contain each topic sorted by the number of grants (largest to smallest) in the Topic network (Grants as edges) constructed using the current data set (2010 to 2016)

Topic	Number of Grants	Value of Grants
criminology	5	£5,080,147.00
development (biosciences)	5	£2,325,834.00
genomics	5	£24,454,128.00
intelligent & expert systems	5	£9,297,777.00
organisational studies	5	£5,952,437.00
power sys man, prot & control	5	£2,609,392.00
digital art & design	4	£18,465,875.00
food processing	4	£7,452,606.00
new media/web-based studies	4	£25,232,837.00
vlsi design	4	£7,404,543.00
assess/remediate contamination	3	£4,534,982.00
climate & climate change	3	£1,750,258.00
criminal law & criminology	3	£5,019,737.00
displays	3	£11,749,150.00
economics	3	£11,015,140.00
electromagnetics	3	£6,188,182.00
food structure/composition	3	£5,141,921.00
macro-molecular delivery	3	£7,026,191.00
media & communication studies	3	£12,545,468.00
pavement engineering	3	£10,890,078.00
population ecology	3	£16,690,970.00
scattering & spectroscopy	3	£7,558,809.00
social policy	3	£1,942,868.00
tools for the biosciences	3	£4,221,581.00
catalysis & enzymology	2	£689,340.00
cells	2	£6,116,997.00
digital arts htp	2	£12,062,042.00
environment & health	2	£568,459.00

Table A.1: Topics and the number and value of grants that contain each topic sorted by the number of grants (largest to smallest) in the Topic network (Grants as edges) constructed using the current data set (2010 to 2016)

Topic	Number of Grants	Value of Grants
governance	2	£1,097,691.00
knowledge management	2	£1,056,671.00
marketing	2	£1,015,702.00
mathematical & statistic psych	2	£1,117,398.00
microeconomic theory	2	£842,950.00
oil & gas extraction	2	£2,642,452.00
plant physiology	2	£1,198,745.00
political geography	2	£7,455,498.00
power electronics	2	£2,270,158.00
power systems plant	2	£3,941,441.00
protein chemistry	2	£389,453.00
regional & extreme weather	2	£2,750,487.00
social anthropology	2	£6,771,898.00
structural biology	2	£524,187.00
ageing: chemistry/biochemistry	1	£475,926.00
animal behaviour	1	£506,361.00
animal organisms	1	£1,581,410.00
applied arts htp	1	£6,358,107.00
behavioural & experimental eco	1	£1,975,496.00
biological membranes	1	£2,215.00
biomedical sciences	1	£188,407.00
carbohydrate chemistry	1	£920,060.00
coal technology	1	£6,550,555.00
composition	1	£5,971,211.00
computational linguistics	1	£340,421.00
computational methods & tools	1	£98,603.00
data handling & storage	1	£742,513.00
development geography	1	£241,076.00

Table A.1: Topics and the number and value of grants that contain each topic sorted by the number of grants (largest to smallest) in the Topic network (Grants as edges) constructed using the current data set (2010 to 2016)

Topic	Number of Grants	Value of Grants
developmental psychology	1	£559,077.00
diamond light source	1	£1,436,518.00
earth & environmental	1	£1,581,410.00
earth engineering	1	£1,000,951.00
education	1	£3,444,605.00
environmental economics	1	£2,926.00
environmental planning	1	£3,567,862.00
evolution & populations	1	£1,190,209.00
geohazards	1	£246,258.00
human geography (general)	1	£3,567,862.00
industrial-org/occupational	1	£3,793,546.00
intelligent measurement sys.	1	£4,956,495.00
international law	1	£1,975,496.00
international relations theory	1	£3,489,111.00
land - ocean interactions	1	£1,415,336.00
macroeconomics	1	£4,000,602.00
mech. & fluid power transmiss.	1	£514,747.00
mental health	1	£294,021.00
microbiology	1	£404,819.00
musical performance	1	£5,971,211.00
novel industrial products	1	£7,174,439.00
plant responses to environment	1	£1,190,209.00
pollution	1	£6,531,178.00
product design	1	£5,965,328.00
protein engineering	1	£1,015,843.00
protein folding / misfolding	1	£458,233.00
research approaches	1	£691,643.00
science & technology studies	1	£2,667,741.00

Table A.1: Topics and the number and value of grants that contain each topic sorted by the number of grants (largest to smallest) in the Topic network (Grants as edges) constructed using the current data set (2010 to 2016)

Topic	Number of Grants	Value of Grants
social stats., comp. & methods	1	£3,444,605.00
social theory	1	£775,542.00
sociology	1	£3,567,862.00
survey & monitoring	1	£246,258.00
systems neuroscience	1	£100,803.00
time-based media htp	1	£6,358,107.00

Table A.2: Statistics of the Topic network (Grants as edges) constructed using the current data set (2010 to 2016). 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.

	uw	**wnn**	**wnv**
Nodes	223	223	223
Edges	2008	2008	2008
Type	Undirected	Undirected	Undirected
Weighted	Yes	Yes	Yes
Connected	Yes	Yes	Yes
Average Degree	18.009	18.009	18.009
Average Weighted Degree	54.897	19.543	22.35
Diameter	8	5	6
Radius	3	3	3
Density	0.081	0.081	0.081
Modularity	0.49	0.373	0.385
Communities	7	6	6
Weak Components	1	1	1
Node Closeness	0.37	0.423	0.412
Node Betweenness	185.088	156.483	162.633
Edge Betweenness	32.882	29.706	30.389
Average Clustering Coefficient	0.597	0.597	0.597
Eigenvector Centrality	0.048	0.204	0.183
Average Path Length	2.395	2.395	2.395

Table A.3: 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 network (Grants as edges) constructed using the current data set (2010 to 2016). 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.

	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

Table A.4: Number of topics clustered within each community discovered as a result of applying several community detection algorithms to the Topic network (Grants as edges) 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. Three of the row names are abbreviated: **uw**, **wnn** and **wnv** stand for unweighted, weighted by normalized number of grants and weighted by normalized value of grants, respectively.

		C1	**C2**	**C3**	**C4**	**C5**	**C6**
uw	**Spinglass**	35	60	25	53	36	14
	Louvain	46	61	36	60	20	-
	Fast Greedy	95	57	60	8	3	-
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	-

Table A.5: Number and value of grants within each community discovered as a result of applying several community detection algorithms to the Topic network (Grants as edges) 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. Six of the column names are abbreviated: **C1** stands for Community 1, **C2** stands for Community 2 and so on. Twelve of the row names are abbreviated: **uw**, **wnn** and **wnv** stand for unweighted, weighted by normalized number of grants and weighted by normalized value of grants, respectively. **SG**, **LV** and **FG** stand for the Spinglass, Louvain and Fast Greedy community detection algorithms.

		C1	C2	C3	C4	C5	C6	Total
uw	**SG**	376	727	483	1402	711	185	3884
		£468M	£830M	£769M	£1.6B	£769M	232M	£4.7B
	LV	590	1351	685	727	583	-	3936
		£689M	£1.7B	£843M	£830M	£729M	-	£4.8B
	FG	1810	1014	773	177	2	-	3776
		£2B	£1.3B	£855M	£294M	1.2M	-	£4.5B
wnn	**SG**	774	453	772	485	1376	-	3860
		£862M	£542M	£892M	£766M	£1.6B	-	£4.7B
	LV	511	774	1338	317	480	484	3904
		£629M	£862M	£1.5B	£332M	£584M	£766M	£4.8B
	FG	718	1680	784	436	-	-	3618
		£801M	£2.2B	£894M	£503M	-	-	£4.4B
wnv	**SG**	1471	386	634	788	498	-	3777
		£1.8B	£381M	£716M	£877M	£607M	-	£4.5B
	LV	1243	317	545	659	565	620	3949
		£1.4B	£332M	£814M	£756M	£662M	£749M	4.8B
	FG	451	1732	887	253	402	-	3725
		£418M	£2B	£1B	£295M	£713M	-	£4.5B

Table A.6: Total number and value of unique grants within communities, total number and value of unique grants within the network and total number and value of unique grants between communities discovered as a result of applying several community detection algorithms to the Topic network (Grants as edges) constructed using the current data set (2010 to 2016). Twelve of the row names are abbreviated: **uw**, **wnn** and **wnv** stand for unweighted, weighted by normalized number of grants and weighted by normalized value of grants, respectively. **SG**, **LV** and **FG** stand for the Spinglass, Louvain and Fast Greedy community detection algorithms.

		Total number and value within communities (unique)	**Total number and value within network (unique)**	**Total number and value between communities (unique)**
uw	**SG**	3059 £3.51B	3103 £3.56B	44 £42M
	LV	3071 £3.53B	3103 £3.56B	32 £27M
	FG	3100 £3.55B	3103 £3.56B	3 £2M
wnn	**SG**	3092 £3.55B	3103 £3.56B	11 £4M
	LV	3072 £3.54B	3103 £3.56B	31 £21M
	FG	3078 £3.53B	3103 £3.56B	25 £26M
wnv	**SG**	3070 £3.53B	3103 £3.56B	33 £22M
	LV	3077 £3.53B	3103 £3.56B	26 £26M
	FG	3070 £3.53B	3103 £3.56B	33 £23M

Table A.7: Topics clustered within each community and sub-community discovered as a result of applying the Louvain community detection algorithm to the Topic network (Grants as edges) constructed using the current data set (2010 to 2016). Six of the row names are abbreviated: **C1** stands for Community 1, **C2** stands 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**	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 A.7: Topics clustered within each community and sub-community discovered as a result of applying the Louvain community detection algorithm to the Topic network (Grants as edges) constructed using the current data set (2010 to 2016). Six of the row names are abbreviated: **C1** stands for Community 1, **C2** stands 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).

	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

Table A.7: Topics clustered within each community and sub-community discovered as a result of applying the Louvain community detection algorithm to the Topic network (Grants as edges) constructed using the current data set (2010 to 2016). Six of the row names are abbreviated: **C1** stands for Community 1, **C2** stands 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).

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
	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

A.1.1.2 Historical data set (2000 to 2010)

Table A.8: Topics and the number and value of grants that contain each topic sorted by the number of grants (largest to smallest) in the Topic network (Grants as edges) constructed using the historical data set (v to 2010)

Topic	Number of Grants	Value of Grants
materials characterisation	2389	£759,734,489.00
materials synthesis & growth	1173	£430,639,300.00
materials processing	899	£320,041,055.00
chemical synthetic methodology	833	£228,631,699.00
fundamentals of computing	673	£137,993,580.00
instrumentation eng. & dev.	569	£223,201,085.00
algebra & geometry	558	£87,645,143.00
catalysis & applied catalysis	534	£128,331,364.00
chemical structure	518	£163,434,000.00
networks & distributed systems	471	£191,491,096.00
image & vision computing	444	£142,156,888.00
surfaces & interfaces	442	£189,026,023.00
information & knowledge mgmt	428	£207,334,615.00
artificial intelligence	424	£113,099,037.00
condensed matter physics	414	£159,648,712.00
materials testing & eng.	411	£179,320,340.00
non-linear systems mathematics	404	£83,766,241.00
mathematical analysis	397	£68,015,950.00
biological & medicinal chem.	392	£156,620,153.00
electronic devices & subsys.	380	£187,642,388.00
chemical biology	377	£146,872,141.00
software engineering	369	£126,787,838.00
numerical analysis	356	£75,426,513.00
analytical science	352	£127,585,114.00
optical devices & subsystems	319	£160,108,893.00
statistics & appl. probability	318	£79,699,402.00
med.instrument.device& equip.	314	£149,512,248.00

Table A.8: Topics and the number and value of grants that contain each topic sorted by the number of grants (largest to smallest) in the Topic network (Grants as edges) constructed using the historical data set (v to 2010)

Topic	Number of Grants	Value of Grants
complex fluids & soft solids	308	£101,368,533.00
lasers & optics	305	£114,747,052.00
continuum mechanics	303	£52,884,668.00
eng. dynamics & tribology	295	£105,820,527.00
human-computer interactions	294	£133,564,251.00
biomaterials	290	£102,615,189.00
medical science & disease	279	£104,492,024.00
mathematical physics	263	£50,724,586.00
transport ops & management	261	£101,968,674.00
gas & solution phase reactions	260	£87,744,036.00
rf & microwave technology	256	£99,319,921.00
control engineering	254	£123,237,305.00
optoelect. devices & circuits	254	£138,710,965.00
digital signal processing	248	£76,784,917.00
system on chip	240	£123,255,458.00
building ops & management	235	£93,514,233.00
combustion	230	£62,256,796.00
aerodynamics	227	£68,273,323.00
fluid dynamics	216	£68,289,650.00
construction ops & management	210	£106,725,995.00
civil engineering materials	204	£67,313,044.00
tissue engineering	203	£75,273,733.00
co-ordination chemistry	202	£52,721,606.00
manufact. enterprise ops& mgmt	190	£97,964,202.00
logic & combinatorics	189	£38,070,447.00
reactor engineering	181	£38,913,260.00
energy efficiency	180	£92,407,766.00
coastal & waterway engineering	179	£41,536,247.00

Table A.8: Topics and the number and value of grants that contain each topic sorted by the number of grants (largest to smallest) in the Topic network (Grants as edges) constructed using the historical data set (v to 2010)

Topic	Number of Grants	Value of Grants
ground engineering	179	£39,464,108.00
multiphase flow	177	£58,497,558.00
water engineering	177	£50,898,445.00
cells	173	£82,995,878.00
solar technology	166	£82,482,682.00
mobile computing	165	£99,965,326.00
acoustics	163	£42,759,760.00
design of process systems	163	£78,410,261.00
mathematical aspects of or	163	£62,084,009.00
new & emerging comp. paradigms	158	£54,258,668.00
quantum optics & information	156	£75,409,081.00
asymmetric chemistry	155	£29,828,358.00
electrochemical science & eng.	154	£49,154,033.00
optical communications	152	£75,655,526.00
vision & senses - ict appl.	151	£30,289,567.00
manufacturing machine & plant	149	£128,618,552.00
design engineering	148	£98,718,525.00
urban & land management	148	£81,998,811.00
particle technology	146	£46,188,181.00
heat & mass transfer	145	£37,599,948.00
microsystems	145	£78,729,303.00
drug formulation & delivery	140	£42,874,611.00
biomedical neuroscience	137	£54,141,902.00
multimedia	136	£47,972,775.00
parallel computing	128	£37,003,920.00
vlsi design	127	£46,009,331.00
magnetism/magnetic phenomena	126	£39,643,014.00
cognitive science appl. in ict	124	£34,034,420.00

Table A.8: Topics and the number and value of grants that contain each topic sorted by the number of grants (largest to smallest) in the Topic network (Grants as edges) constructed using the historical data set (v to 2010)

Topic	Number of Grants	Value of Grants
design & testing technology	123	£88,150,086.00
scattering & spectroscopy	122	£30,584,497.00
bioprocess engineering	121	£30,620,921.00
design processes	121	£104,501,536.00
fuel cell technologies	121	£46,342,449.00
human communication in ict	121	£20,144,796.00
structural engineering	120	£23,242,706.00
cold atomic species	118	£48,847,399.00
bioinformatics	115	£42,302,496.00
manufact. business strategy	115	£110,712,471.00
separation processes	108	£24,116,911.00
theoretical biology	107	£56,546,768.00
oil & gas extraction	106	£20,587,376.00
modelling & simul. of it sys.	102	£35,914,067.00
plasmas - laser & fusion	102	£37,262,964.00
high performance computing	100	£19,145,394.00
physical organic chemistry	100	£33,071,090.00
power sys man, prot & control	96	£43,354,612.00
optical phenomena	95	£62,703,041.00
genomics	94	£57,085,247.00
light-matter interactions	94	£30,240,221.00
computer graphics & visual.	92	£24,505,256.00
assess/remediate contamination	91	£25,224,774.00
robotics & autonomy	89	£29,667,115.00
nuclear structure	86	£23,145,374.00
combinatorial chemistry	79	£43,819,067.00
electric motor & drive systems	79	£31,754,971.00
rheology	79	£32,175,984.00

Table A.8: Topics and the number and value of grants that contain each topic sorted by the number of grants (largest to smallest) in the Topic network (Grants as edges) constructed using the historical data set (v to 2010)

Topic	Number of Grants	Value of Grants
complexity science	78	£27,358,020.00
bioelectronic devices	73	£59,252,437.00
power electronics	73	£46,397,828.00
sustainable energy vectors	73	£47,686,076.00
bioenergy	70	£37,946,111.00
plasmas - technological	69	£12,785,855.00
development (biosciences)	67	£17,763,030.00
population ecology	65	£23,685,236.00
sustainable energy networks	64	£76,321,710.00
power systems plant	63	£36,516,930.00
waste minimisation	63	£18,223,472.00
energy - conventional	62	£44,460,426.00
intelligent & expert systems	61	£36,013,579.00
waste management	59	£31,585,718.00
biomechanics & rehabilitation	57	£31,226,401.00
quantum fluids & solids	57	£19,577,127.00
energy storage	55	£33,428,341.00
electromagnetics	52	£25,946,802.00
displays	50	£38,143,712.00
comput./corpus linguistics	45	£12,458,975.00
intelligent measurement sys.	45	£13,409,588.00
medical imaging	43	£36,199,261.00
escience	41	£38,022,845.00
energy - nuclear	39	£42,728,955.00
mining & minerals extraction	39	£7,152,469.00
wind power	39	£23,671,142.00
carbohydrate chemistry	36	£21,765,823.00
energy - marine & hydropower	36	£24,951,360.00

Table A.8: Topics and the number and value of grants that contain each topic sorted by the number of grants (largest to smallest) in the Topic network (Grants as edges) constructed using the historical data set (v to 2010)

Topic	Number of Grants	Value of Grants
pavement engineering	33	£15,717,787.00
coal technology	31	£21,285,362.00
music & acoustic technology	29	£5,449,813.00
fusion	26	£164,347,711.00
mech. & fluid power transmiss.	23	£6,319,201.00
biophysics	21	£7,431,980.00
safety & reliability of plant	18	£10,624,681.00
animal & human physiology	15	£1,540,368.00
computer sys. & architecture	15	£11,141,810.00
catalysis & enzymology	14	£6,138,451.00
carbon capture & storage	12	£19,049,408.00
synthetic biology	12	£10,252,952.00
education	11	£2,501,690.00
new media/web-based studies	10	£25,175,136.00
biological membranes	9	£1,459,588.00
media & communication studies	9	£17,215,909.00
protein chemistry	9	£967,564.00
management & business studies	7	£1,871,987.00
musculoskeletal system	7	£635,908.00
underwater engineering	6	£3,127,427.00
atoms & ions	5	£1,680,041.00
psychology	5	£14,038,727.00
design htp	4	£12,422,292.00
economics	4	£24,202,309.00
psycholinguistics	4	£443,798.00
publishing	4	£1,213,731.00
applied arts htp	3	£105,995.00
astron. & space sci. technol.	3	£224,705.00

Table A.8: Topics and the number and value of grants that contain each topic sorted by the number of grants (largest to smallest) in the Topic network (Grants as edges) constructed using the historical data set (v to 2010)

Topic	Number of Grants	Value of Grants
digital art & design	3	£12,381,437.00
digital arts htp	3	£14,889,857.00
food processing	3	£1,209,056.00
language acquisition	3	£192,896.00
product design	3	£12,380,275.00
protein folding / misfolding	3	£466,251.00
structural biology	3	£822,929.00
bionanoscience	2	£477,392.00
bionanotechnology	2	£125,500.00
galactic & interstellar astron	2	£142,346.00
interpreting & translation	2	£329,655.00
language training/educational	2	£121,859.00
policy, arts mgmt & creat ind	2	£2,084,929.00
pollution	2	£427,121.00
sociology	2	£2,910,846.00
tools for the biosciences	2	£133,480.00
accelerator r&d	1	£34,914.00
agricultural systems	1	£11,814,897.00
applied linguistics	1	£62,765.00
archaeology of literate soc.	1	£29,505.00
cell cycle	1	£118,280.00
crop science	1	£1,302,692.00
cultural history	1	£2,708,997.00
cultural studies & pop culture	1	£12,584,584.00
drama & theatre - other	1	£38,565.00
economic & social history	1	£2,708,997.00
environmental informatics	1	£623,007.00
environmental planning	1	£11,814,897.00

Table A.8: Topics and the number and value of grants that contain each topic sorted by the number of grants (largest to smallest) in the Topic network (Grants as edges) constructed using the historical data set (v to 2010)

Topic	Number of Grants	Value of Grants
evolution & populations	1	£142,910.00
food structure/composition	1	£144,592.00
human geography	1	£11,814,897.00
languages & linguistics	1	£47,417.00
mantle & core processes	1	£903,958.00
mental health	1	£12,100,779.00
microbiology	1	£63,956.00
science-based archaeology	1	£61,168.00
social stats., comp. & methods	1	£488,208.00
sociolinguistics	1	£62,765.00
soil science	1	£1,302,692.00
stem cell biology	1	£101,745.00
upper atmos process & geospace	1	£219,973.00

Table A.9: Statistics of the Topic network (Grants as edges) constructed using the historical data set (2000 to 2010). 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.

	uw	**wnn**	**wnv**
Nodes	208	208	208
Edges	3592	3592	3592
Type	Undirected	Undirected	Undirected
Weighted	Yes	Yes	Yes
Connected	No	No	No
Average Degree	34.538	34.538	34.538
Average Weighted Degree	222.933	35.337	36.423
Diameter	7	5	5
Radius	1	1	1
Density	0.167	0.167	0.167
Modularity	0.436	0.271	0.279
Communities	7	5	5
Weak Components	2	2	2
Node Closeness	0.221	0.245	0.244
Node Betweenness	145.778	109.776	110.5
Edge Betweenness	14.32	12.235	12.277
Average Clustering Coefficient	0.59	0.59	0.59
Eigenvector Centrality	0.031	0.232	0.227
Average Path Length	2.077	2.077	2.077

Table A.10: 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 network (Grants as edges) constructed using the historical data set (2000 to 2010). 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.

	uw		**wnn**		**wnv**	
Infomap	6	0.007	4	0.002	4	0.002
Spinglass	-	-	-	-	-	-
Louvain	5	0.264	5	0.271	5	0.279
Label Propagation	2	0.001	2	0.001	2	0.001
Leading Eigenvector	5	0.243	7	0.248	7	0.254
Walktrap	24	0.177	31	0.205	29	0.229
Fast Greedy	5	0.222	5	0.251	5	0.262
Edge Betweenness	123	0.042	121	0.044	101	0.04

Table A.11: Number of topics clustered within each community discovered as a result of applying several community detection algorithms to the Topic network (Grants as edges) constructed using the historical data set (2000 to 2010). Five of the column names are abbreviated: **C1** stands for Community 1, **C2** stands for Community 2 and so on. Three of the row names are abbreviated: **uw**, **wnn** and **wnv** stand for unweighted, weighted by normalized number of grants and weighted by normalized value of grants, respectively.

		C1	**C2**	**C3**	**C4**	**C5**
uw	**Spinglass**	-	-	-	-	-
	Louvain	67	43	25	2	71
	Fast Greedy	93	22	89	2	2
wnn	**Spinglass**	-	-	-	-	-
	Louvain	67	43	25	2	71
	Fast Greedy	89	46	69	2	2
wnv	**Spinglass**	-	-	-	-	-
	Louvain	30	67	2	58	51
	Fast Greedy	63	51	17	75	2

Table A.12: Number and value of grants within each community discovered as a result of applying several community detection algorithms to the Topic network (Grants as edges) constructed using the historical data set (2000 to 2010). 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. Five of the column names are abbreviated: **C1** stands for Community 1, **C2** stands for Community 2 and so on. Twelve of the row names are abbreviated: **uw**, **wnn** and **wnv** stand for unweighted, weighted by normalized number of grants and weighted by normalized value of grants, respectively. **SG**, **LV** and **FG** stand for the Spinglass, Louvain and Fast Greedy community detection algorithms.

		C1	C2	C3	C4	C5	Total
uw	**SG**	-	-	-	-	-	-
		-	-	-	-	-	-
	LV	8682	5167	1394	1	4099	19343
		£2.6B	£1.3B	£699M	£1.3M	£1.3B	£6B
	FG	6459	2987	9863	2	1	19312
		£2B	£993M	£3B	£133M	£1.3M	£6.1B
wnn	**SG**	-	-	-	-	-	-
		-	-	-	-	-	-
	LV	8682	5167	1394	1	4099	19343
		£2.6B	£1.3B	£699M	£1.3M	£1.3B	£6B
	FG	6373	3782	9087	2	1	19245
		£1.7B	£1.4B	£2.8B	£133K	£1.3M	£6B
wnv	**SG**	-	-	-	-	-	-
		-	-	-	-	-	-
	LV	3393	8798	1	2857	4785	19834
		£808M	£2.6B	£1.3M	£972M	£1.7B	£6.2B
	FG	3234	4532	2043	9648	1	19458
		£1B	£1.6B	£408M	£2.9B	£1.3M	£6B

Table A.13: Total number and value of unique grants within communities, total number and value of unique grants within the network and total number and value of unique grants between communities discovered as a result of applying several community detection algorithms to the Topic network (Grants as edges) constructed using the historical data set (2000 to 2010). Twelve of the row names are abbreviated: **uw**, **wnn** and **wnv** stand for unweighted, weighted by normalized number of grants and weighted by normalized value of grants, respectively. **SG**, **LV** and **FG** stand for the Spinglass, Louvain and Fast Greedy community detection algorithms.

		Total number and value within communities (unique)	**Total number and value within network (unique)**	**Total number and value between communities (unique)**
uw	**SG**	-	16768	-
		-	£4.9B	-
	LV	16617	16768	151
		£4.9B	£4.9B	£23M
	FG	16716	16768	52
		£4.9B	£4.9B	£9M
wnn	**SG**	-	16768	-
		-	£4.9B	-
	LV	16617	16768	151
		4.9B	4.9B	£23M
	FG	16722	16768	46
		£4.9	£4.9B	£7M
wnv	**SG**	-	16768	-
		-	£4.9B	-
	LV	16617	16768	151
		£4.9B	£4.9B	£23M
	FG	16614	16768	154
		£4.9M	£4.9M	£24M

Table A.14: Topics clustered within each community and sub-community discovered as a result of applying the Louvain community detection algorithm to the Topic network (Grants as edges) constructed using the historical data set (2000 to 2010). Six of the row names are abbreviated: **C1** stands for Community 1, **C2** stands 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 A.14: Topics clustered within each community and sub-community discovered as a result of applying the Louvain community detection algorithm to the Topic network (Grants as edges) constructed using the historical data set (2000 to 2010). Six of the row names are abbreviated: **C1** stands for Community 1, **C2** stands 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).

	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 A.14: Topics clustered within each community and sub-community discovered as a result of applying the Louvain community detection algorithm to the Topic network (Grants as edges) constructed using the historical data set (2000 to 2010). Six of the row names are abbreviated: **C1** stands for Community 1, **C2** stands 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).

	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

A.1.1.3 Historical data set (1990 to 2000)

Table A.15: Topics and the number and value of grants that contain each topic sorted by the number of grants (largest to smallest) in the Topic network (Grants as edges) constructed using the historical data set (1990 to 2000)

Topic	Number of Grants	Value of Grants
materials characterisation	1437	£250,687,834.00
materials processing	748	£203,166,413.00
materials synthesis & growth	628	£145,266,958.00
instrumentation eng. & dev.	479	£86,796,767.00
chemical synthetic methodology	470	£51,412,325.00
condensed matter physics	372	£95,053,530.00
civil engineering materials	294	£25,079,586.00
catalysis & applied catalysis	288	£41,348,790.00
software engineering	279	£35,721,229.00
surfaces & interfaces	272	£64,912,950.00
algebra & geometry	268	£14,900,406.00
eng. dynamics & tribology	268	£29,145,433.00
optoelect. devices & circuits	264	£105,673,619.00
chemical structure	253	£40,453,029.00
combustion	230	£25,241,206.00
networks & distributed systems	228	£28,072,961.00
image & vision computing	227	£31,233,662.00
atoms & ions	224	£46,364,062.00
mathematical analysis	212	£10,058,459.00
biological & medicinal chem.	211	£25,235,606.00
design & testing technology	203	£41,945,562.00
design engineering	201	£36,814,905.00
fundamentals of computing	197	£21,206,545.00
non-linear systems mathematics	196	£14,789,767.00
building ops & management	190	£17,788,250.00
materials testing & eng.	188	£28,185,972.00

Table A.15: Topics and the number and value of grants that contain each topic sorted by the number of grants (largest to smallest) in the Topic network (Grants as edges) constructed using the historical data set (1990 to 2000)

Topic	Number of Grants	Value of Grants
asymmetric chemistry	187	£12,313,075.00
transport ops & management	185	£16,336,183.00
rf & microwave technology	184	£31,149,929.00
lasers & optics	183	£40,701,092.00
digital signal processing	182	£22,004,380.00
complex fluids & soft solids	181	£25,507,581.00
information & knowledge mgmt	167	£20,543,912.00
statistics & appl. probability	167	£9,415,713.00
manufact. business strategy	161	£20,137,296.00
aerodynamics	157	£18,337,451.00
construction ops & management	155	£16,333,438.00
separation processes	150	£15,390,746.00
continuum mechanics	149	£9,277,276.00
control engineering	144	£17,376,770.00
parallel computing	143	£19,450,475.00
optical devices & subsystems	141	£32,936,520.00
numerical analysis	137	£7,901,228.00
analytical science	135	£22,979,378.00
particle technology	135	£16,776,482.00
coastal & waterway engineering	134	£13,981,192.00
manufact. enterprise ops& mgmt	134	£19,340,531.00
artificial intelligence	133	£16,834,515.00
ground engineering	133	£11,986,760.00
electronic devices & subsys.	131	£66,679,316.00
electrochemical science & eng.	129	£15,001,993.00
multiphase flow	128	£14,152,281.00
manufacturing machine & plant	125	£31,421,846.00

Table A.15: Topics and the number and value of grants that contain each topic sorted by the number of grants (largest to smallest) in the Topic network (Grants as edges) constructed using the historical data set (1990 to 2000)

Topic	Number of Grants	Value of Grants
water engineering	125	£12,927,527.00
design of process systems	121	£34,608,994.00
fluid dynamics	121	£14,296,939.00
human communication in ict	117	£16,907,508.00
gas & solution phase reactions	116	£18,223,275.00
human-computer interactions	116	£13,686,888.00
magnetism/magnetic phenomena	116	£10,388,905.00
optical communications	113	£26,182,578.00
waste minimisation	108	£13,960,979.00
biomaterials	106	£28,736,961.00
nuclear structure	105	£56,485,878.00
design processes	102	£23,411,382.00
oil & gas extraction	93	£11,019,413.00
electric motor & drive systems	92	£12,038,073.00
mathematical physics	90	£5,859,517.00
energy efficiency	80	£9,621,048.00
urban & land management	75	£9,099,271.00
heat & mass transfer	73	£7,483,859.00
co-ordination chemistry	68	£6,228,805.00
intelligent measurement sys.	68	£8,550,439.00
robotics & autonomy	65	£8,514,063.00
solar technology	65	£8,448,834.00
plasmas - laser & fusion	59	£12,421,790.00
logic & combinatorics	58	£1,774,069.00
quantum fluids & solids	57	£9,503,058.00
multimedia	56	£7,560,581.00
plasmas - technological	56	£6,505,639.00

Table A.15: Topics and the number and value of grants that contain each topic sorted by the number of grants (largest to smallest) in the Topic network (Grants as edges) constructed using the historical data set (1990 to 2000)

Topic	Number of Grants	Value of Grants
rheology	50	£5,332,002.00
intelligent & expert systems	49	£6,395,361.00
vlsi design	49	£14,850,916.00
cognitive science appl. in ict	48	£4,200,498.00
physical organic chemistry	46	£4,441,567.00
fuel cell technologies	44	£5,698,208.00
power sys man, prot & control	44	£5,110,041.00
assess/remediate contamination	42	£4,953,719.00
waste management	42	£4,825,299.00
power electronics	41	£5,666,961.00
power systems plant	39	£4,760,209.00
mathematical aspects of or	37	£5,061,582.00
optical phenomena	34	£4,493,384.00
safety & reliability of plant	33	£3,330,300.00
system on chip	33	£6,091,931.00
vision & senses - ict appl.	33	£3,475,973.00
mech. & fluid power transmiss.	32	£6,860,993.00
combinatorial chemistry	30	£31,979,080.00
displays	29	£4,170,045.00
energy storage	28	£3,824,556.00
electromagnetics	25	£3,082,744.00
pavement engineering	25	£2,628,005.00
underwater engineering	25	£2,526,036.00
carbohydrate chemistry	24	£1,920,505.00
coal technology	23	£1,844,856.00
reactor engineering	23	£2,635,800.00
computer graphics & visual.	21	£3,943,981.00

Table A.15: Topics and the number and value of grants that contain each topic sorted by the number of grants (largest to smallest) in the Topic network (Grants as edges) constructed using the historical data set (1990 to 2000)

Topic	Number of Grants	Value of Grants
wind power	17	£1,563,905.00
mining & minerals extraction	14	£1,292,033.00
med.instrument.device& equip.	12	£2,352,888.00
acoustics	8	£1,492,933.00
energy - conventional	8	£1,599,229.00
quantum optics & information	8	£1,372,472.00
scattering & spectroscopy	7	£1,743,101.00
theoretical biology	7	£2,621,923.00
microsystems	6	£732,237.00
tissue engineering	6	£4,297,807.00
bioenergy	5	£562,963.00
energy - nuclear	4	£348,634.00
medical science & disease	4	£256,576.00
sustainable energy networks	4	£505,989.00
cold atomic species	3	£480,395.00
development (biosciences)	3	£156,308.00
music & acoustic technology	3	£535,230.00
sustainable energy vectors	3	£250,025.00
bioprocess engineering	2	£893,176.00
cells	2	£727,467.00
light-matter interactions	2	£173,047.00
bioelectronic devices	1	£49,679.00
catalysis & enzymology	1	£104,794.00
chemical biology	1	£104,794.00
drug formulation & delivery	1	£51,697.00
mobile computing	1	£1,500,000.00
modelling & simul. of it sys.	1	£51,943.00

Table A.15: Topics and the number and value of grants that contain each topic sorted by the number of grants (largest to smallest) in the Topic network (Grants as edges) constructed using the historical data set (1990 to 2000)

Topic	Number of Grants	Value of Grants
population ecology	1	£52,668.00
tools for the biosciences	1	£131,372.00

Table A.16: Statistics of the Topic network (Grants as edges) constructed using the historical data set (1990 to 2000). 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.

	uw	**wnn**	**wnv**
Nodes	136	136	136
Edges	748	748	748
Type	Undirected	Undirected	Undirected
Weighted	Yes	Yes	Yes
Connected	Yes	Yes	Yes
Average Degree	11	11	11
Average Weighted Degree	37.221	12.721	12.618
Diameter	12	6	6
Radius	3	3	3
Density	0.081	0.081	0.081
Modularity	0.461	0.4	0.424
Communities	6	5	6
Weak Components	1	1	1
Node Closeness	0.309	0.392	0.398
Node Betweenness	135.211	108.536	106.257
Edge Betweenness	36.856	32.007	31.592
Average Clustering Coefficient	0.453	0.453	0.453
Eigenvector Centrality	0.046	0.105	0.084
Average Path Length	2.54	2.54	2.54

Table A.17: 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 network (Grants as edges) constructed using the historical data set (1990 to 2000). 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.

	uw		**wnn**		**wnv**	
Infomap	10	0.346	10	0.368	10	0.419
Spinglass	7	0.396	6	0.41	7	0.426
Louvain	7	0.387	5	0.4	6	0.424
Label Propagation	1	0.0	3	0.096	1	0.0
Leading Eigenvector	5	0.371	8	0.349	8	0.373
Walktrap	20	0.346	23	0.371	20	0.395
Fast Greedy	6	0.377	6	0.407	5	0.414
Edge Betweenness	49	0.238	26	0.293	25	0.246

Table A.18: Number of topics clustered within each community discovered as a result of applying several community detection algorithms to the Topic network (Grants as edges) constructed using the historical data set (1990 to 2000). Six of the column names are abbreviated: **C1** stands for Community 1, **C2** stands for Community 2 and so on. Three of the row names are abbreviated: **uw**, **wnn** and **wnv** stand for unweighted, weighted by normalized number of grants and weighted by normalized value of grants, respectively.

		C1	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**
uw	**Spinglass**	40	17	3	3	2	33	38
	Louvain	17	23	29	21	5	30	11
	Fast Greedy	22	48	22	39	3	2	-
wnn	**Spinglass**	34	8	28	17	10	39	-
	Louvain	28	25	17	39	27	-	-
	Fast Greedy	19	34	41	29	5	8	-
wnv	**Spinglass**	1	7	40	42	9	17	20
	Louvain	27	17	37	24	10	21	-
	Fast Greedy	19	27	47	37	6	-	-

Table A.19: Number and value of grants within each community discovered as a result of applying several community detection algorithms to the Topic network (Grants as edges) constructed using the historical data set (1990 to 2000). 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. Six of the column names are abbreviated: **C1** stands for Community 1, **C2** stands for Community 2 and so on. Twelve of the row names are abbreviated: **uw**, **wnn** and **wnv** stand for unweighted, weighted by normalized number of grants and weighted by normalized value of grants, respectively. **SG**, **LV** and **FG** stand for the Spinglass, Louvain and Fast Greedy community detection algorithms.

		C1	C2	C3	C4	C5	C6	C7	Total
uw	**SG**	3894	1328	18	29	8	4342	3781	13400
		£551M	£88M	£6M	£3.1M	£1.3M	£702M	£512M	£1.8B
	LV	1328	1485	3434	3275	57	2707	1107	13393
		£88M	£250M	£538M	£569M	£11M	£334M	£202M	£1.9B
	FG	1706	6602	1574	3282	18	8	-	13190
		£137M	£979M	£299M	£448M	£6.6M	£1.3M	-	£1.8B
wnn	**SG**	3203	514	4156	1328	152	3762	-	13115
		£443M	£64M	£682M	£88M	£23M	£535M	-	£1.8B
	LV	2015	3995	1328	3455	2567	-	-	13360
		£246M	£661M	£88M	£496M	£385M	-	-	£1.9B
	FG	1480	3230	3560	4407	57	514	-	13248
		£105M	£453M	£484M	£730M	£11M	64M	-	£1.8B
wnv	**SG**	0	249	3757	3909	144	1328	3574	12961
		£0	49M	465M	552M	21M	88M	590M	£1.7B
	LV	1820	1328	3362	2520	184	3609	-	12823
		£227M	£88M	£471M	£392M	£26M	£594M	-	£1.8B
	FG	1480	4347	4128	3384	98	-	-	13437
		£105M	£721M	£508M	£515M	£16M	-	-	£1.9B

Table A.20: Total number and value of unique grants within communities, total number and value of unique grants within the network and total number and value of unique grants between communities discovered as a result of applying several community detection algorithms to the Topic network (Grants as edges) constructed using the historical data set (1990 to 2000). Twelve of the row names are abbreviated: **uw**, **wnn** and **wnv** stand for unweighted, weighted by normalized number of grants and weighted by normalized value of grants, respectively. **SG**, **LV** and **FG** stand for the Spinglass, Louvain and Fast Greedy community detection algorithms.

		Total number and value within communities (unique)	**Total number and value within network (unique)**	**Total number and value between communities (unique)**
uw	**SG**	12899	13179	280
		£1.8B	£1.7B	£27M
	LV	12563	13179	616
		£1.7B	£1.7B	£64M
	FG	12740	13179	439
		£1.7B	£1.7B	£44M
wnn	**SG**	12595	13179	584
		£1.7B	£1.7B	£60M
	LV	12791	13179	388
		£1.7B	£1.7B	£36M
	FG	12704	13179	475
		£1.7B	£1.7B	£48M
wnv	**SG**	12473	13179	706
		£1.6B	£1.7B	£122M
	LV	12274	13179	905
		£1.6B	£1.7B	£99M
	FG	12871	13179	308
		£1.8B	£1.7B	£31M

Table A.21: Topics clustered within each community and sub-community discovered as a result of applying the Louvain community detection algorithm to the Topic network (Grants as edges) constructed using the historical data set (1990 to 2000). Six of the row names are abbreviated: **C1** stands for Community 1, **C2** stands 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 A.21: Topics clustered within each community and sub-community discovered as a result of applying the Louvain community detection algorithm to the Topic network (Grants as edges) constructed using the historical data set (1990 to 2000). Six of the row names are abbreviated: **C1** stands for Community 1, **C2** stands 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).

	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

B Word cloud representations

B.1 Networks of Topics (Grants as edges)

B.1.1 Based on word frequency

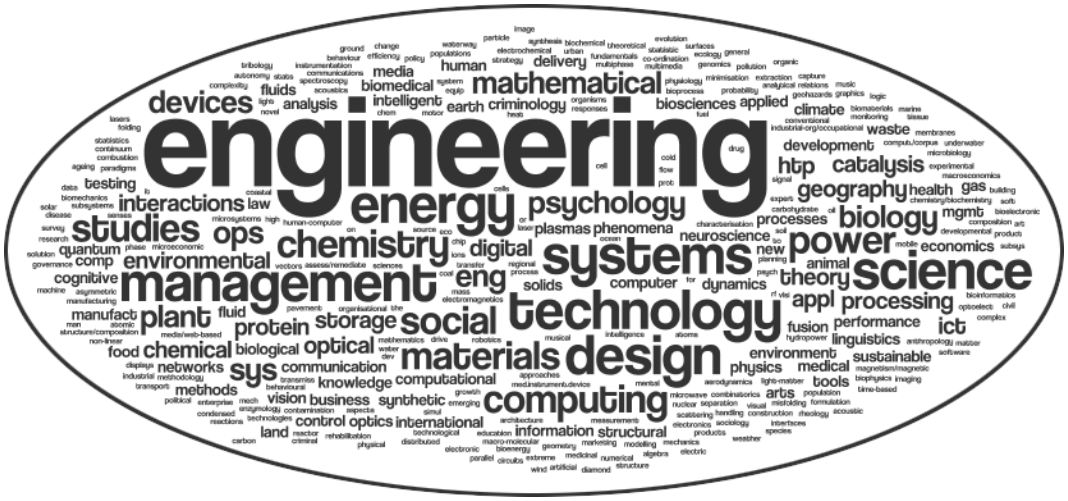


Figure B.1: Word cloud representation created using Wordle showcasing words that formulate the topics found in the Topic network (Grants as edges). Font size represents of word frequency within the the text corpus.

B.1.2 Based on the number of grants containing topics

Figure B.2: Word cloud representation created using Wordle showcasing topics found in the Topic network (Grants as edges). Font size represents the number of grants containing a specific topic.

B.1.3 Based on the value of grants containing topics

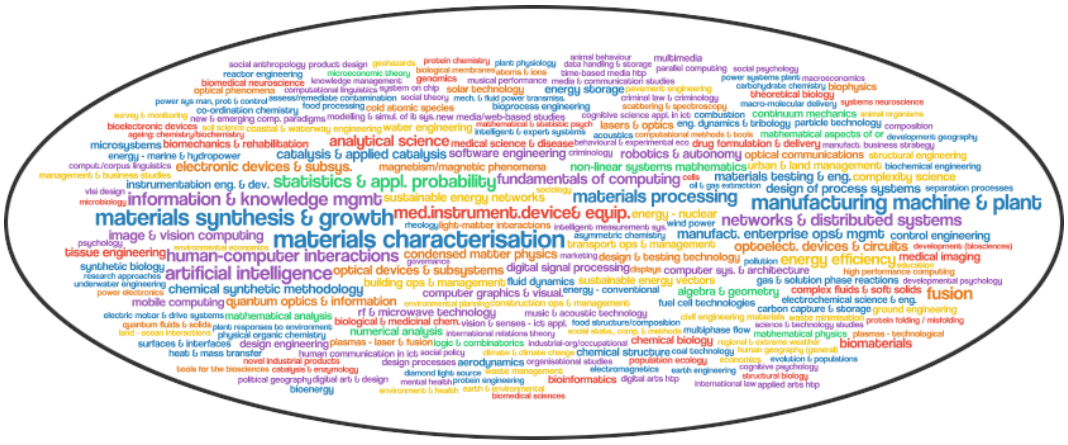


Figure B.3: Word cloud representation created using Wordle showcasing topics topics found in the Topic network (Grants as edges). Font size represents the number of grants containing a specific topic.

B.2 Communities of topics

B.2.1 Community 1

B.2.1.1 Based on word frequency

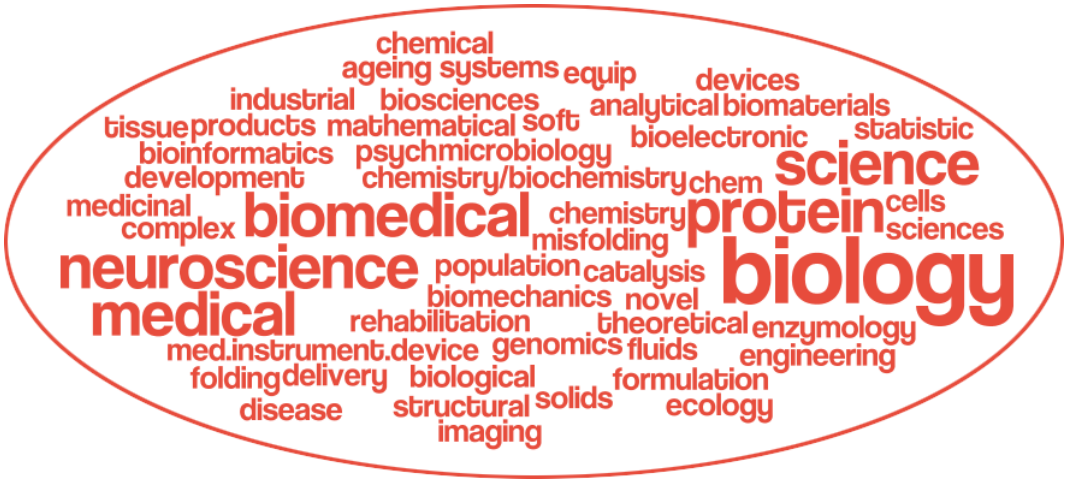


Figure B.4: Word cloud representation created using Wordle showcasing words that formulate the topics clustered within Community 1 as identified by the Louvain community detection algorithm. Font size represents the frequency of the word in the text corpus made out of all the words that formulate the topics clustered in Community 1.

B.2.1.2 Based on the number of grants containing topics

Figure B.5: Word cloud representation created using Wordle showcasing topics clustered within Community 1 as identified by the Louvain community detection algorithm. Font size represents the number of grants containing a specific topic.

B.2.1.3 Based on the value of grants containing topics

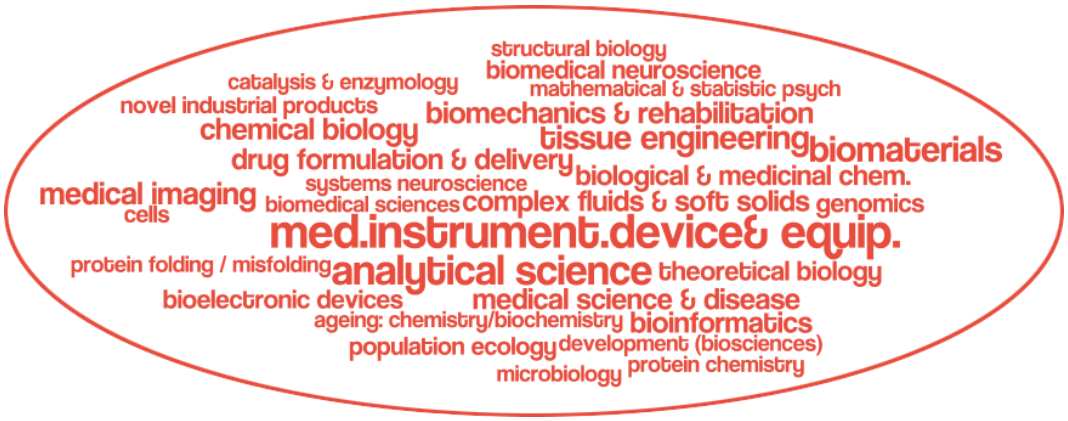


Figure B.6: Word cloud representation created using Wordle showcasing topics clustered within Community 1 as identified by the Louvain community detection algorithm. Font size represents the number of grants containing a specific topic.

B.2.2 Community 2

B.2.2.1 Based on word frequency

Figure B.7: Word cloud representation created using Wordle showcasing words that formulate the topics clustered within Community 2 as identified by the Louvain community detection algorithm. Font size represents the frequency of the word in the text corpus made out of all the words that formulate the topics clustered in Community 2.

B.2.2.2 Based on the number of grants containing topics

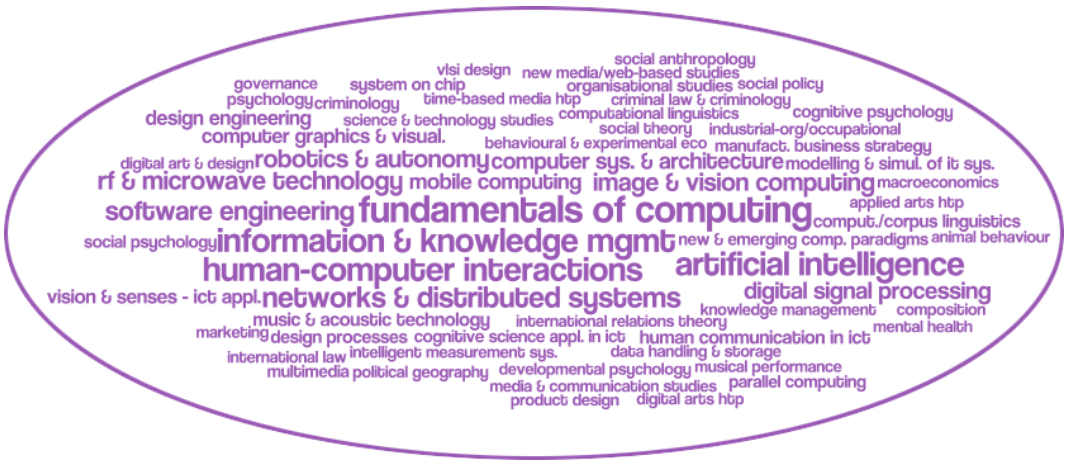


Figure B.8: Word cloud representation created using Wordle showcasing topics clustered within Community 2 as identified by the Louvain community detection algorithm. Font size represents the number of grants containing a specific topic.

B.2.2.3 Based on the value of grants containing topics

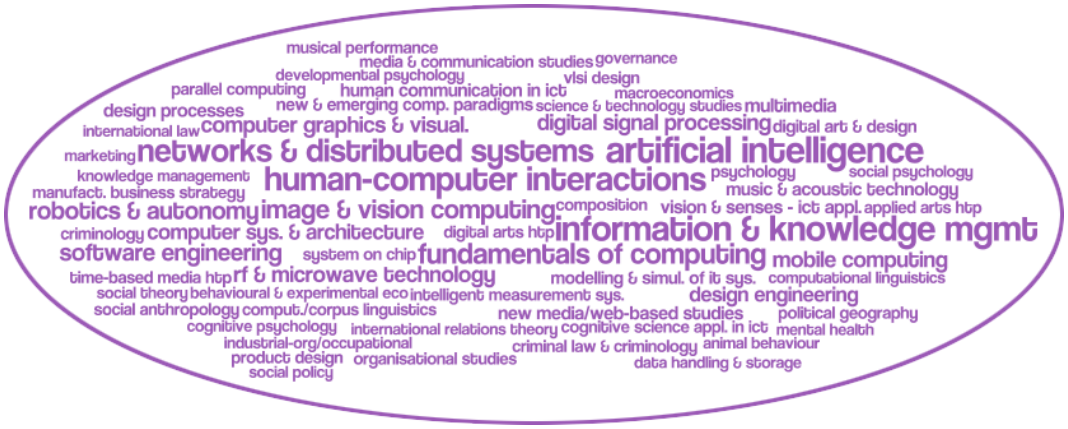


Figure B.9: Word cloud representation created using Wordle showcasing topics clustered within Community 2 as identified by the Louvain community detection algorithm. Font size represents the number of grants containing a specific topic.

B.2.3 Community 3

B.2.3.1 Based on word frequency

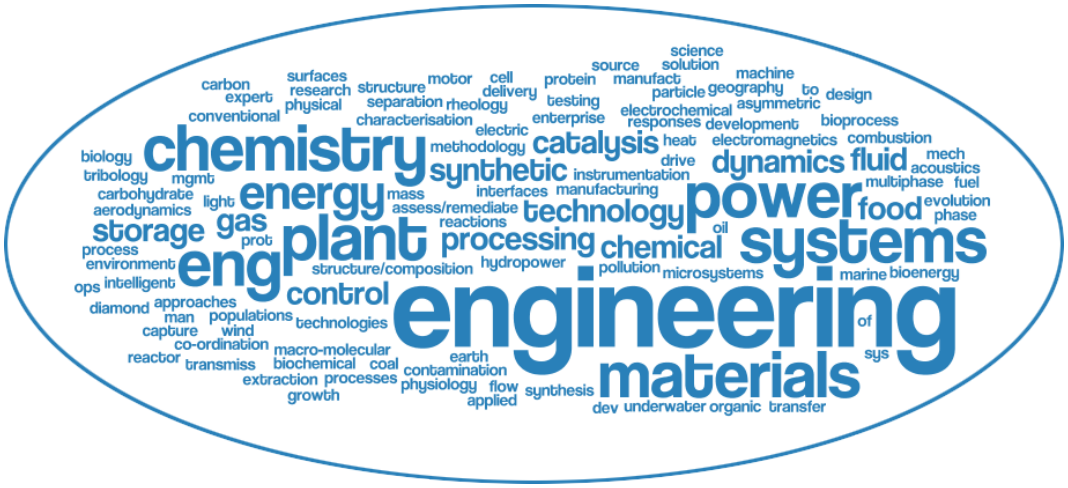


Figure B.10: Word cloud representation created using Wordle showcasing words that formulate the topics clustered within Community 1 as identified by the Louvain community detection algorithm. Font size represents the frequency of the word in the text corpus made out of all the words that formulate the topics clustered in Community 3.

B.2.3.2 Based on the number of grants containing topics

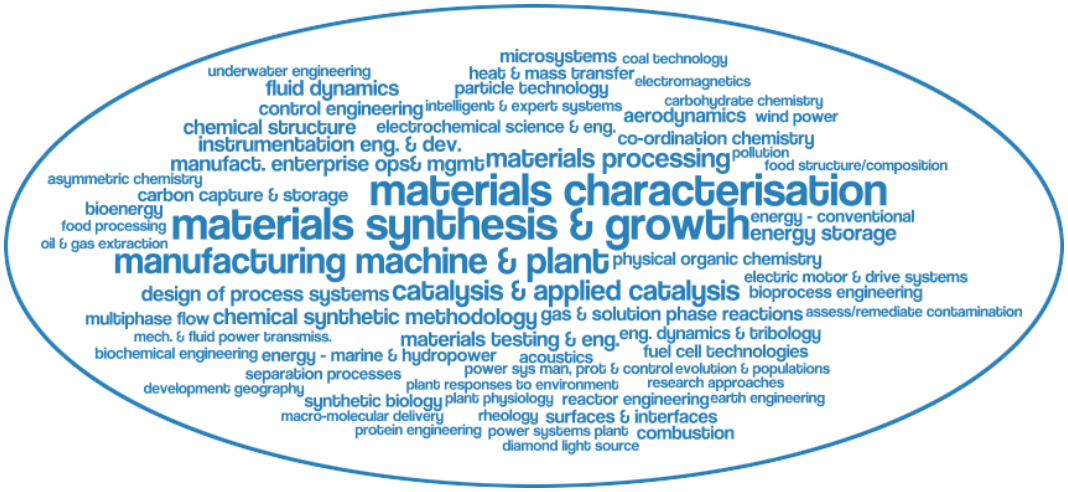


Figure B.11: Word cloud representation created using Wordle showcasing topics clustered within Community 3 as identified by the Louvain community detection algorithm. Font size represents the number of grants containing a specific topic.

B.2.3.3 Based on the value of grants containing topics

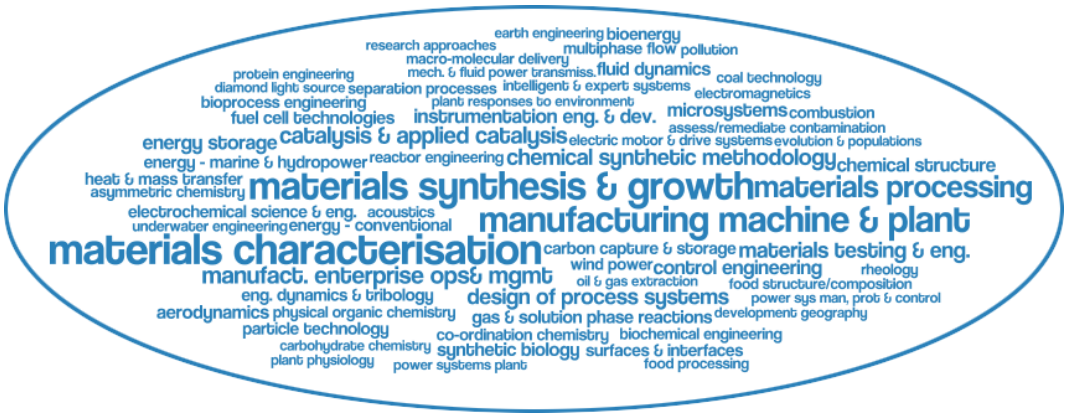


Figure B.12: Word cloud representation created using Wordle showcasing topics clustered within Community 3 as identified by the Louvain community detection algorithm. Font size represents the number of grants containing a specific topic.

B.2.4 Community 4

B.2.4.1 Based on word frequency

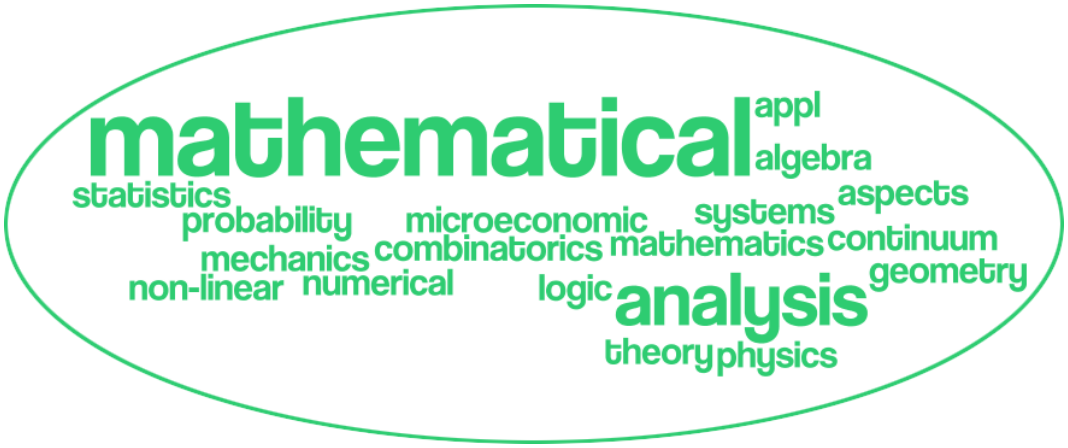


Figure B.13: Word cloud representation created using Wordle showcasing words that formulate the topics clustered within Community 4 as identified by the Louvain community detection algorithm. Font size represents the frequency of the word in the text corpus made out of all the words that formulate the topics clustered in Community 4.

B.2.4.2 Based on the number of grants containing topics

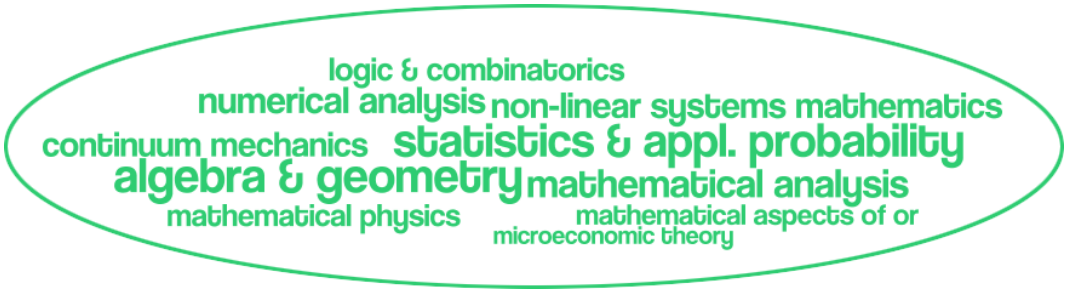


Figure B.14: Word cloud representation created using Wordle showcasing topics clustered within Community 4 as identified by the Louvain community detection algorithm. Font size represents the number of grants containing a specific topic.

B.2.4.3 Based on the value of grants containing topics

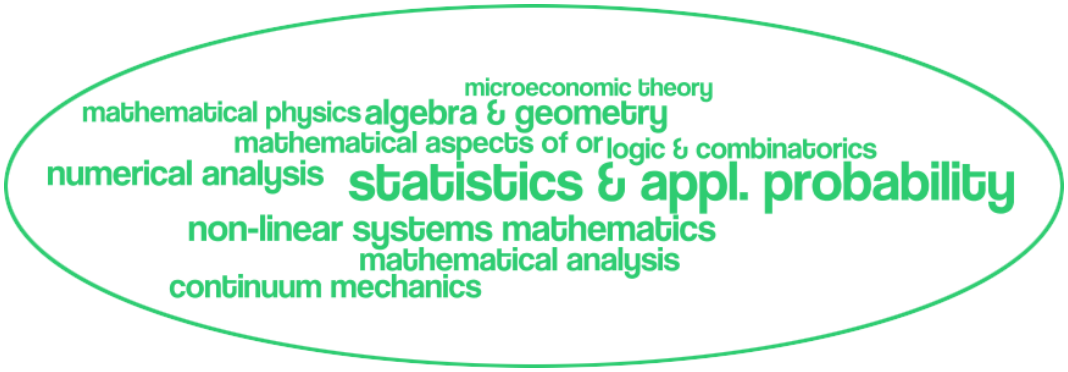


Figure B.15: Word cloud representation created using Wordle showcasing topics clustered within Community 4 as identified by the Louvain community detection algorithm. Font size represents the number of grants containing a specific topic.

B.2.5 Community 5

B.2.5.1 Based on word frequency

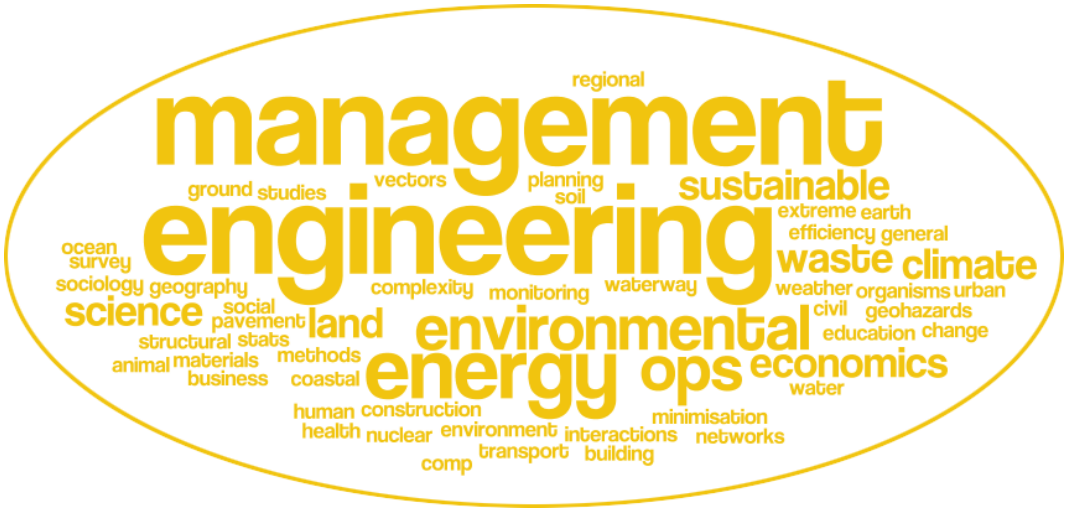


Figure B.16: Word cloud representation created using Wordle showcasing words that formulate the topics clustered within Community 5 as identified by the Louvain community detection algorithm. Font size represents the frequency of the word in the text corpus made out of all the words that formulate the topics clustered in Community 5.

B.2.5.2 Based on the number of grants containing topics

Figure B.17: Word cloud representation created using Wordle showcasing topics clustered within Community 5 as identified by the Louvain community detection algorithm. Font size represents the number of grants containing a specific topic.

B.2.5.3 Based on the value of grants containing topics

Figure B.18: Word cloud representation created using Wordle showcasing topics clustered within Community 5 as identified by the Louvain community detection algorithm. Font size represents the number of grants containing a specific topic.

B.2.6 Community 6

B.2.6.1 Based on word frequency

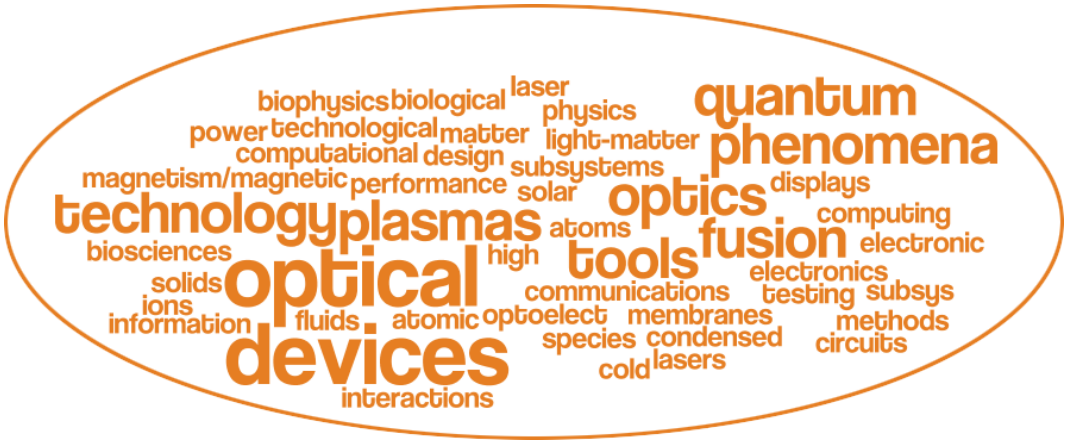


Figure B.19: Word cloud representation created using Wordle showcasing words that formulate the topics clustered within Community 6 as identified by the Louvain community detection algorithm. Font size represents the frequency of the word in the text corpus made out of all the words that formulate the topics clustered in Community 6.

B.2.6.2 Based on the number of grants containing topics

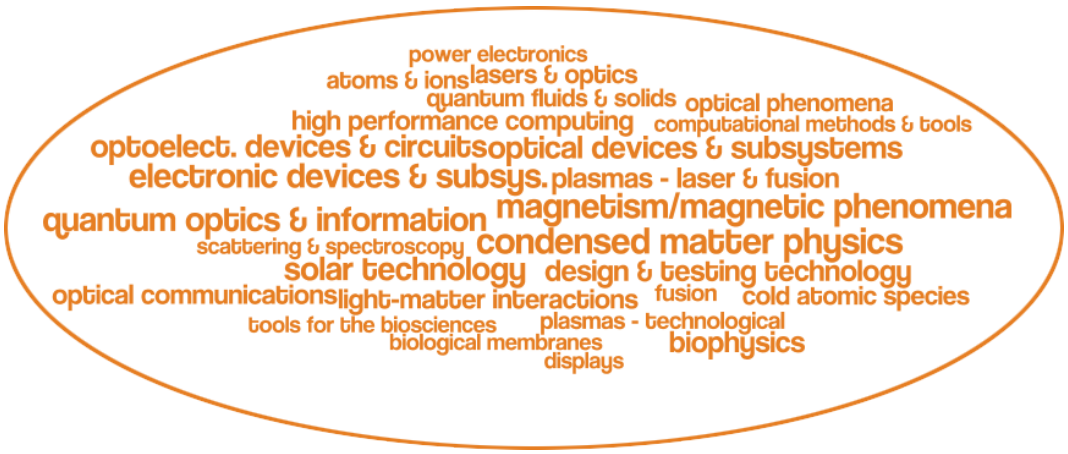


Figure B.20: Word cloud representation created using Wordle showcasing topics clustered within Community 6 as identified by the Louvain community detection algorithm. Font size represents the number of grants containing a specific topic.

B.2.6.3 Based on the value of grants containing topics

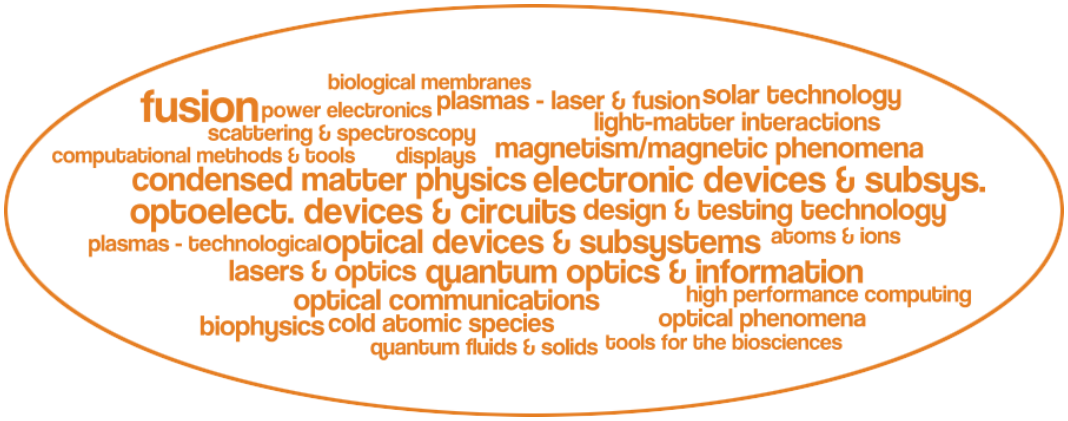


Figure B.21: Word cloud representation created using Wordle showcasing topics clustered within Community 6 as identified by the Louvain community detection algorithm. Font size represents the number of grants containing a specific topic.

B.3 Sub-communities of topics

B.3.1 Sub-communities within Community 1

B.3.1.1 Based on the number of grants containing topics

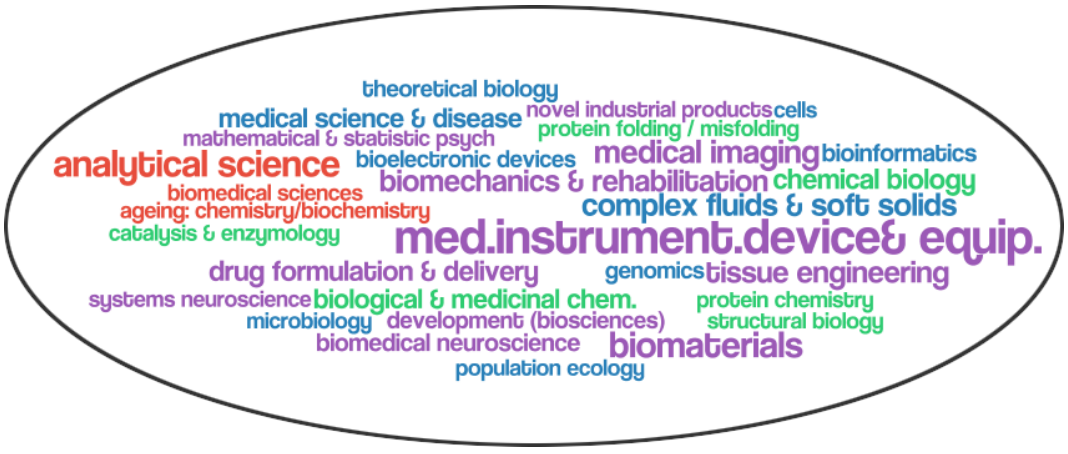


Figure B.22: Word cloud representation showcasing topics in sub-communities within Community 1 as identified by the Louvain community detection algorithm. Font size represents the number of grants containing a specific topic, while text colour represents the sub-communities identified within Community 1.

B.3.1.2 Based on the value of grants containing topics

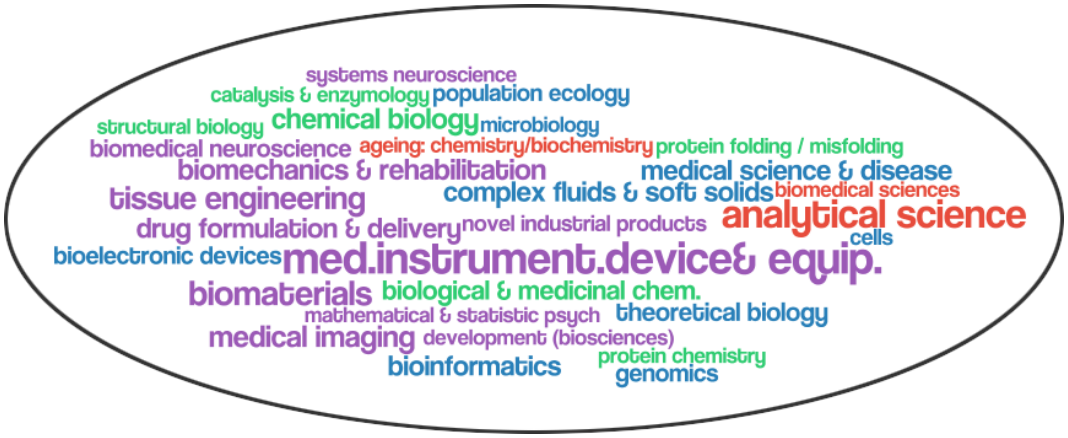


Figure B.23: Word cloud representation showcasing topics in sub-communities within Community 1 as identified by the Louvain community detection algorithm. Font size represents the value of grants containing a specific topic, while text colour represents the sub-communities identified within Community 1.

B.3.2 Sub-communities within Community 2

B.3.2.1 Based on the number of grants containing topics

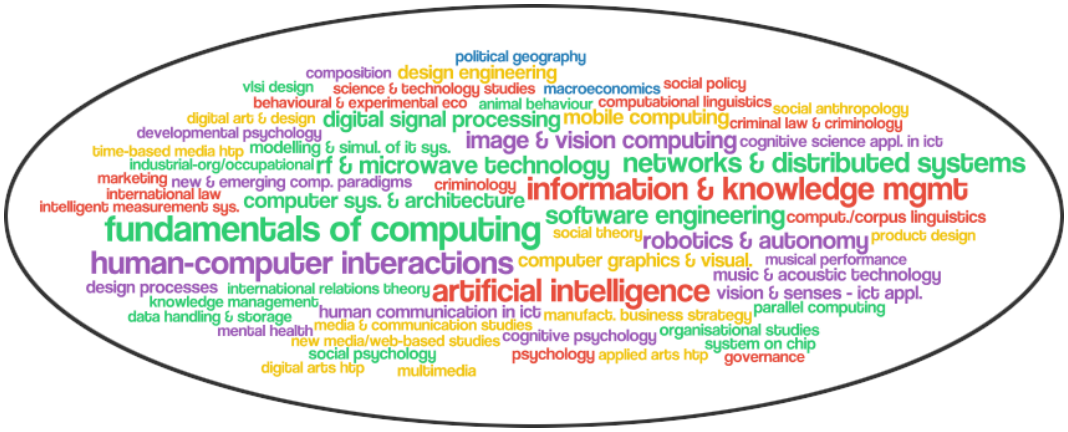


Figure B.24: Word cloud representation showcasing topics in sub-communities within Community 2 as identified by the Louvain community detection algorithm. Font size represents the number of grants containing a specific topic, while text colour represents the sub-communities identified within Community 2.

B.3.2.2 Based on the value of grants containing topics

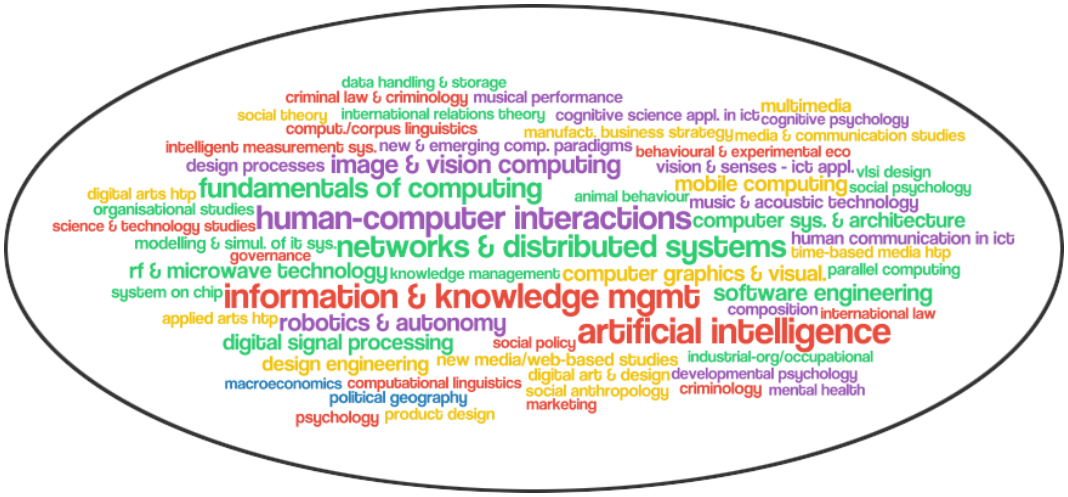


Figure B.25: Word cloud representation showcasing topics in sub-communities within Community 2 as identified by the Louvain community detection algorithm. Font size represents the value of grants containing a specific topic, while text colour represents the sub-communities identified within Community 2.

B.3.3 Sub-communities within Community 3

B.3.3.1 Based on the number of grants containing topics

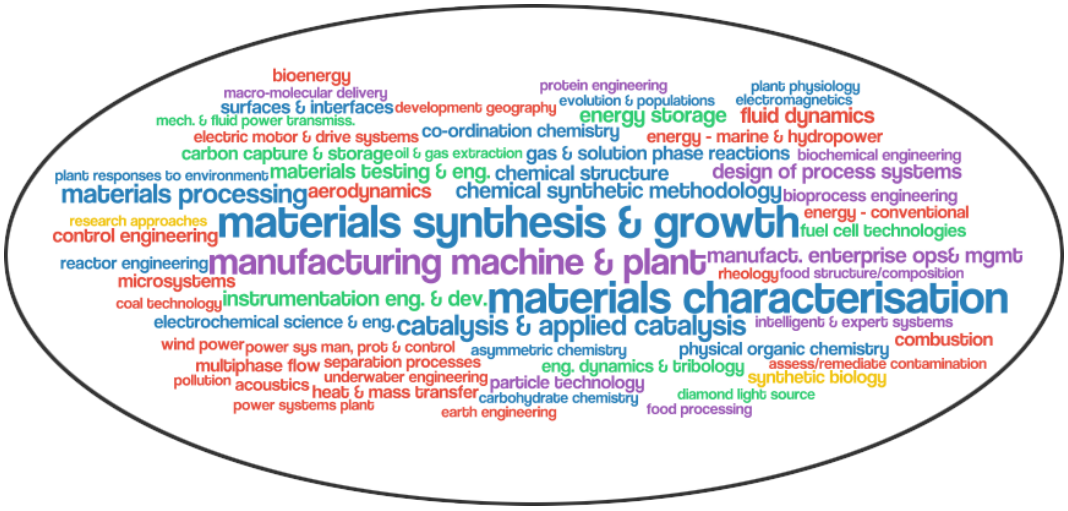


Figure B.26: Word cloud representation showcasing topics in sub-communities within Community 3 as identified by the Louvain community detection algorithm. Font size represents the number of grants containing a specific topic, while text colour represents the sub-communities identified within Community 3.

B.3.3.2 Based on the value of grants containing topics

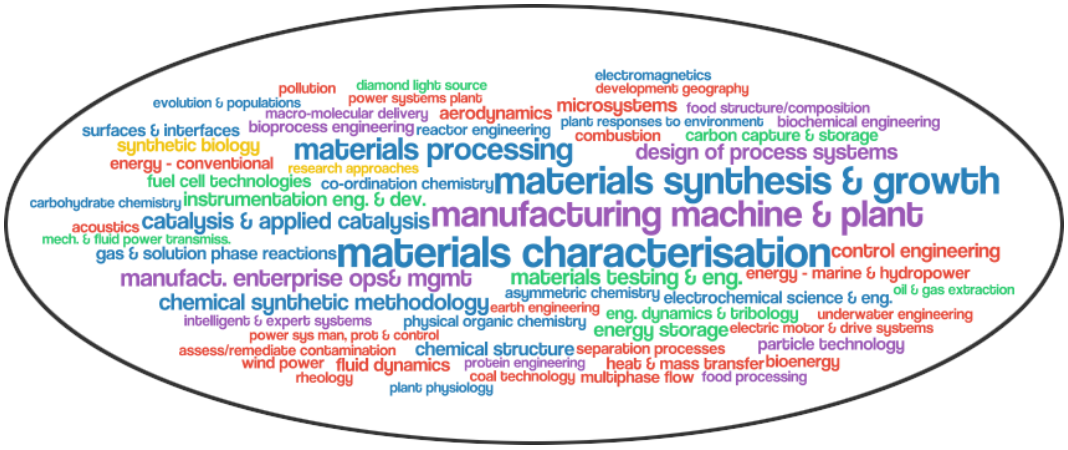


Figure B.27: Word cloud representation showcasing topics in sub-communities within Community 3 as identified by the Louvain community detection algorithm. Font size represents the value of grants containing a specific topic, while text colour represents the sub-communities identified within Community 3.

B.3.4 Sub-communities within Community 4

B.3.4.1 Based on the number of grants containing topics

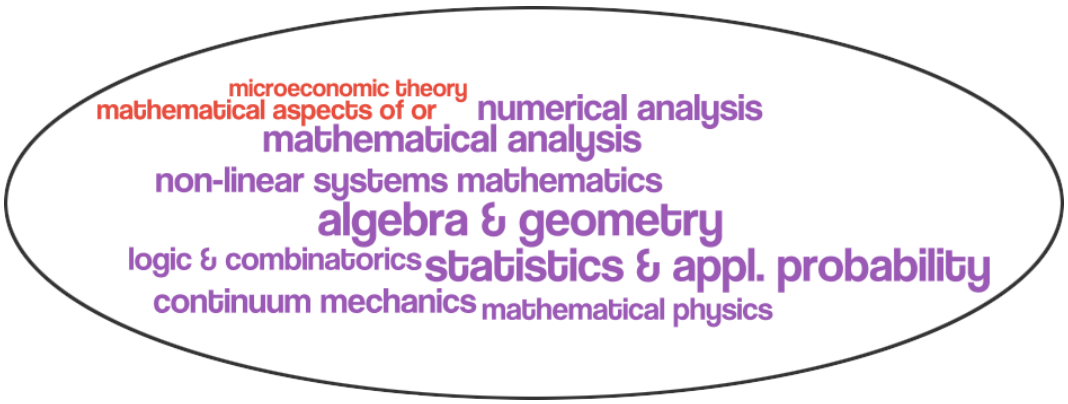


Figure B.28: Word cloud representation showcasing topics in sub-communities within Community 4 as identified by the Louvain community detection algorithm. Font size represents the number of grants containing a specific topic, while text colour represents the sub-communities identified within Community 4.

B.3.4.2 Based on the value of grants containing topics

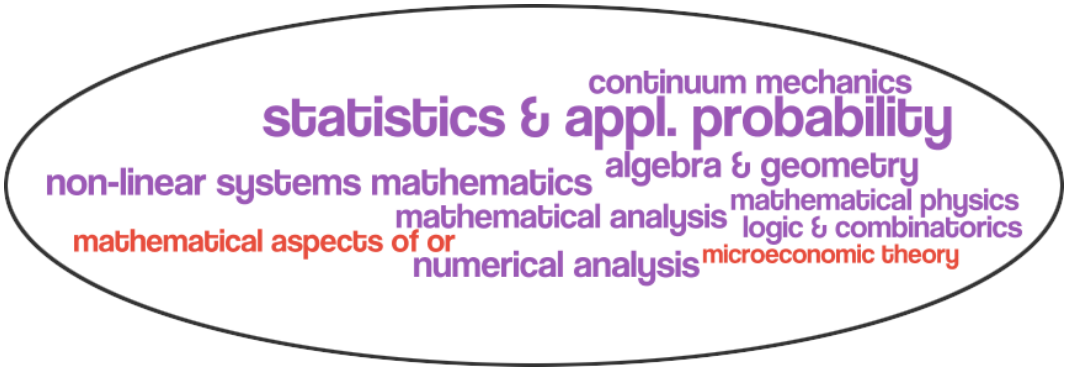


Figure B.29: Word cloud representation showcasing topics in sub-communities within Community 4 as identified by the Louvain community detection algorithm. Font size represents the value of grants containing a specific topic, while text colour represents the sub-communities identified within Community 4.

B.3.5 Sub-communities within Community 5

B.3.5.1 Based on the number of grants containing topics

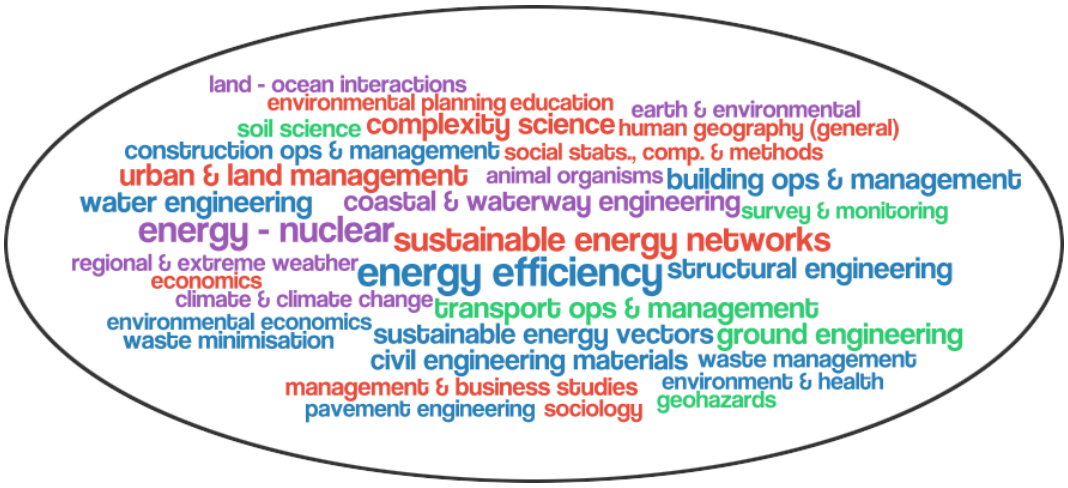


Figure B.30: Word cloud representation showcasing topics in sub-communities within Community 5 as identified by the Louvain community detection algorithm. Font size represents the number of grants containing a specific topic, while text colour represents the sub-communities identified within Community 5.

B.3.5.2 Based on the value of grants containing topics

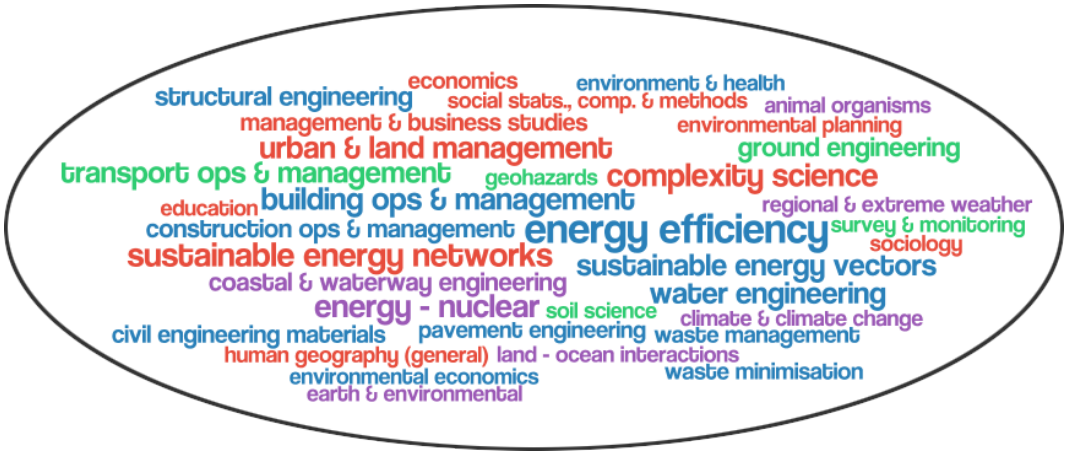


Figure B.31: Word cloud representation showcasing topics in sub-communities within Community 5 as identified by the Louvain community detection algorithm. Font size represents the value of grants containing a specific topic, while text colour represents the sub-communities identified within Community 5.

B.3.6 Sub-communities within Community 6

B.3.6.1 Based on the number of grants containing topics

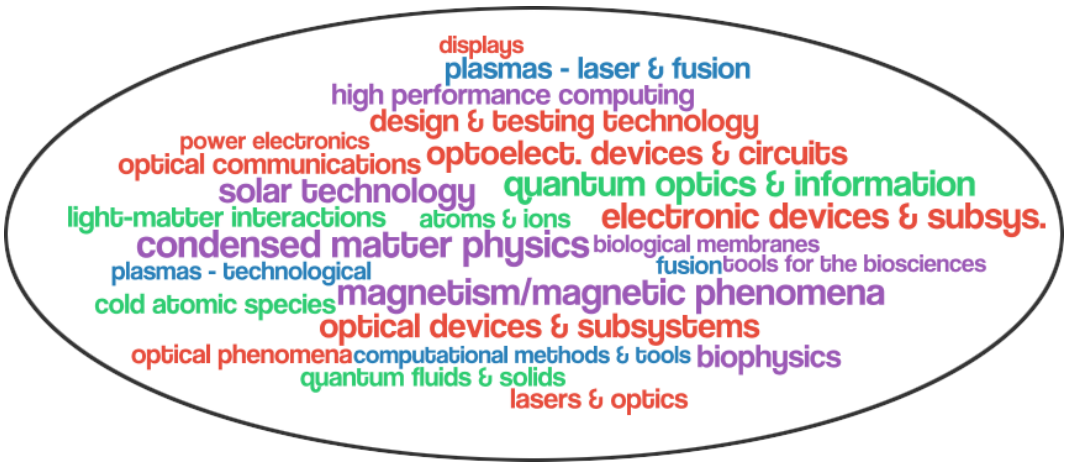


Figure B.32: Word cloud representation showcasing topics in sub-communities within Community 6 as identified by the Louvain community detection algorithm. Font size represents the number of grants containing a specific topic, while text colour represents the sub-communities identified within Community 6.

B.3.6.2 Based on the value of grants containing topics

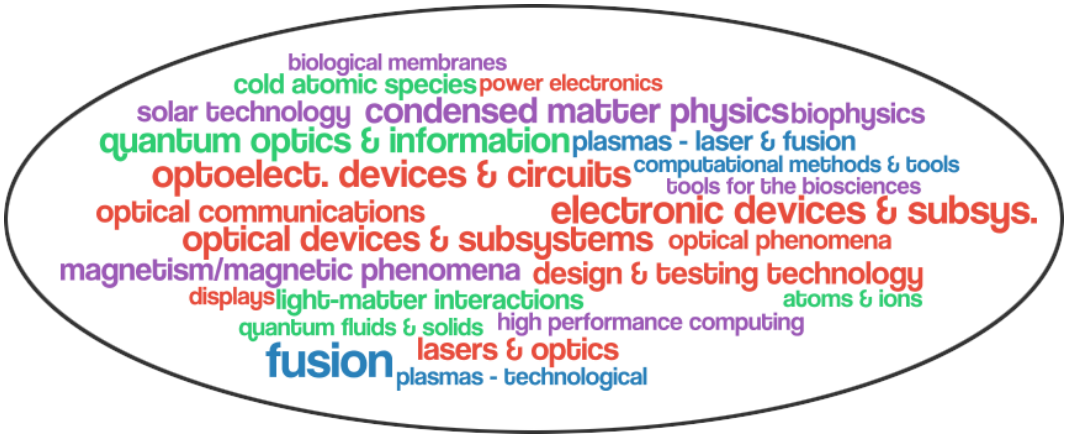


Figure B.33: Word cloud representation showcasing topics in sub-communities within Community 6 as identified by the Louvain community detection algorithm. Font size represents the value of grants containing a specific topic, while text colour represents the sub-communities identified within Community 6.

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] 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.

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

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

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

[8] Wikipedia. Community structure. URL: https://en.wikipedia.org/wiki/Community_structure (visited on 29/07/2016).

[9] Lei Tang and Huan Liu. Community detection and mining in social media. URL: http://dmml.asu.edu/cdm/ (visited on 29/07/2016).

[10] Lei Tang and Huan Liu. Community detection and mining in social media. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 2(1):1–137, 2010.

[11] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[12] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[13] Ali Reihanian, Behrouz Minaei-Bidgoli, and Hosein Alizadeh. Topic-oriented community detection of rating-based social networks. *Journal of King Saud University-Computer and Information Sciences*, 2015.

[14] Zhongying Zhao, Shengzhong Feng, Qiang Wang, Joshua Zhexue Huang, Graham J Williams, and Jianping Fan. Topic oriented community detection through social objects and link analysis in social networks. *Knowledge-Based Systems*, 26:164–173, 2012.

[15] Andrea Lancichinetti and Santo Fortunato. Community detection algorithms: a comparative analysis. *Physical review E*, 80(5):056117, 2009.

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

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

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

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

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

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