Quantifying textual similarities across scientific research communities

Boyd Duffee

Doctor of Philosophy

School of Computer Science and Mathematics, Keele University.

June 2018

Quantifying textual similarities across scientific research communities

Boyd Duffee

Doctor of Philosophy

School of Computer Science and Mathematics, Keele University.

June 2018

Contents

1 Introduction **3**
1.1 Networks 4
1.2 Text Mining 5
1.3 The Relation to Knowledge 5
1.4 Research Questions 6
1.5 Summary 7
2 Data Acquisition and Curation **8**
2.1 Motivation 8
2.2 Text Corpus 8
2.2.1 Selection Criteria 9
2.3 Citation Network 13
2.4 Document Processing 16
2.5 Summary 18
3 Citation Networks **19**
3.1 Introduction to Networks 19
3.1.1 Common Features of Complex Networks 24
3.1.2 Clustering 25
3.1.3 Formal Definition of Complex Networks 27
3.2 Acquiring the Data 28
3.2.1 Selection Criteria 29
3.2.2 Data Collection 30
3.2.3 Methods 31
3.3 Community 32
3.3.1 Properties 36
3.4 Community Detection Algorithms 38
3.5 Modelling Citation Networks 41
3.6 Analysis of Dark Matter Citation Network 42

3.7 Summary . . . 46

4 Text Mining . . . 48

4.1 Introduction . . . 48

4.1.1 Knowledge Representation . . . 52

4.1.2 The Vector Space Model for Documents . . . 53

4.2 Document Selection . . . 54

4.3 Parsing . . . 55

4.3.1 Tokenising . . . 55

4.3.2 Stop Words . . . 56

4.3.3 Normalisation . . . 57

4.3.4 Stemming . . . 57

4.3.5 Discussion on Parsing . . . 58

4.4 Computing Similarity using the Vector Space Model . . . 59

4.4.1 Using Cosine as a Measure of Similarity . . . 59

4.4.2 Comparison of Vector Space Model variants . . . 59

4.5 Method . . . 61

4.6 Clustering . . . 63

4.7 Results . . . 65

4.7.1 Distributions . . . 66

4.7.2 Similarity over Time . . . 67

4.7.3 Residual Sum of Squares . . . 70

4.8 Cross-validating Louvain Clustering with k-means . . . 73

4.9 Cluster Labelling . . . 79

4.10 Summary . . . 82

5 Observations on Communities . . . 84

5.1 Categorisation of the Dark Matter corpus . . . 84

5.2 k-core visualisation . . . 85

5.3 Correlations . . . 88

5.4 Adjusted Rand Index . . . 91

5.5 Vector Space Model as a network . . . 93
5.6 Community Identity . . . 95
5.7 Summary . . . 102

6 General Discussion . . . 103
6.1 Networks . . . 104
6.1.1 Citations and Graphs . . . 104
6.1.2 Technical Challenges . . . 105
6.1.3 Selection Criteria . . . 107
6.1.4 Network Clustering . . . 107
6.1.5 Pair Correlation . . . 108
6.1.6 Growth of Networks . . . 110
6.1.7 Significance . . . 110
6.2 Text Mining . . . 111
6.2.1 Text Clustering . . . 113
6.3 Communities . . . 115
6.3.1 Document Graph . . . 120
6.4 Summary . . . 121

7 Conclusion . . . 122
7.1 Further Work . . . 123
7.1.1 Testing Models . . . 123
7.1.2 Document Analysis . . . 123
7.1.3 Terminological Distance . . . 124
7.1.4 Bipartite Graphs . . . 125
7.1.5 Topic Modelling and Citation Communities . . . 126

A Repositories . . . 128
A.1 ADS - the SAO/NASA Astrophysics Data System . . . 128
A.2 arXiv . . . 129

B LaTeXML . . . 131
B.1 LaTeXML and Equations . . . 132

C Algorithms and Software **133**
D Mathematical Notes **139**
E The Dark Matter Problem **142**
F Stop Words and Group Data **145**
Glossary **163**

References

List of Figures

2.1 arXiv search results for Dark Matter in 2009 by year . . . 10
2.2 Web of Science search results for Dark Matter . . . 11
2.3 Estimated coverage of arXiv . . . 12
2.4 Shell 1 – Citations to and References from the Dark Matter core . . . 14
2.5 Shell 2 – Citations to and References from nodes in Shell 1 . . . 15
3.1 A diagram of the Königsberg Bridge Problem . . . 20
3.2 Network diagram example: 2 nodes . . . 22
3.3 Network diagram example: 3 nodes . . . 22
3.4 Clustering distributions by network shells . . . 32
3.5 Clustering distributions by network shells—log-scale plot . . . 33
3.6 Clustering coefficient as a function of node degree, k . . . 34
3.7 Distribution of Community Sizes . . . 37
3.8 In-degree distribution of the citation network . . . 43
3.9 CDF of theIn-degree distribution . . . 44
3.10 Pair correlation of citations . . . 45
3.11 Pair correlation below $k = 30$ . . . 46
4.1 Text Similarity distribution between all documents . . . 67
4.2 Text Similarity over Time . . . 68
4.3 Text Similarity distributions over Time . . . 69
4.4 Residual Sum of Squares $\widehat{RSS}_{min}$ for k-means clustering . . . 70
4.5 Comparison of $\widehat{RSS}_{min}$ with a straight line . . . 71
4.6 Applying Louvain clustering to document text similarity network . . . 73
4.7 Comparison of text clusters found with k-means and Louvain . . . 74
4.8 Comparison of set similarities among thresholded partitions . . . 76
4.9 Comparison of set similarities among partitions produced by k-means . . . 77
4.10 Comparison of set similarities between partitions with different K . . . 78
4.11 Monte Carlo simulation with document similarity network . . . 79
4.12 Monte Carlo simulation ΔSimilarity/Random . . . 80
5.1 k-core visualisation of the Citation Network core papers . . . 86
5.2 k-core visualisation of the VSM model . . . 87
5.3 Correlation of text similarity with citation network community . . . 89
5.4 Correlations of similarities between and within citation communities . . . 91
5.5 Adjusted Rand Index of Network communities with k-means clusters . . . 92
5.6 Adjusted Rand Index of Network communities with k-means clusters . . . 93
5.7 Probability distribution for document graphs with different thresholds . . . 94
5.8 Document graph degree distribution, $\theta = 0.6$ . . . 95
5.9 Visualisation of communities . . . 97

List of Tables

1.1 A taxonomy of areas of study 4
2.1 File types of Papers in the Dark Matter corpus received from arXiv . . 17
3.1 Network Statistics 42
4.1 Percentile values of the number of unique terms in each document . . . 63
5.1 Categories of Papers in the Dark Matter corpus 85
5.2 General network statistics for the document graph 96
F.1 Lingua::EN::StopWords 145
F.2 Extended stopword list derived from the Dark Matter Corpus 146
F.3 Constituent Data for Group Membership in the citation network clustering 147
F.4 Group comparisons of Similarity, Group Size and Link Saturation . . . 149
F.5 arXiv ids constituting the Dark Matter corpus 154

Abstract

There are well-established approaches of text mining collections of documents and for understanding the network of citations between academic papers. Few studies have examined the textual content of the papers that constitute a citation network. A document corpus was obtained from the arXiv repository, selected from papers relating to the subject of Dark Matter and a citation network was created from the data held by NASA's Astrophysics Data System on those papers, their citations and references.

I use the Louvain community-finding algorithm on the Dark Matter network to identify groups of papers with a higher density of citations and compare the textual similarity between papers in the Dark Matter corpus using the Vector Space Model of document representation and the cosine similarity function.

It was found that pairs of papers within a citation community have a higher similarity than they do with papers in other citation communities. This implies that content is associated with structure in scientific citation networks, which opens avenues for research on network communities for finding ground-truth using advanced Text Mining techniques, such as Topic Modelling. It was found that using the titles of papers in a citation network community was a good method for identifying the community. The power law exponent of the degree distribution was found to be, $\gamma = 2.3$, lower than results reported for other citation networks. The selection of papers based on a single subject, rather than based on a journal or category, is suggested as the reason for this lower value. It was also found that the degree pair correlation of the citation network classifies it as a disassortative network with a cut-off value at degree $k_c = 30$. The textual similarity of documents decreases linearly with age over a 15 year time span.

Acknowledgements

This research has made use of the arXiv open-access repository and NASA's Astrophysics Data System. For providing the text documents, I thank the arXiv project and arXiv administrator, Jake Weiskoff, for gathering all my requested files and providing them as a single tar file download. For providing access to their citation data, I thank the ADS and Program Manager, Alberto Accomazzi, who provided advice on downloading issues.

I acknowledge funding from the School of Computing and Mathematics and the Directorate of Finance & IT, both at Keele University for the partial payment of my tuition fees. I thank my supervisor, Dr. Gordon Rugg, for his support, patience and advice over many long years.

I apologise to my wife and daughter for too many years of being "half there" and thank them for their humour and understanding. Finally, I thank my older brother, Neil, for never giving me the answer, but instead giving me the skills to work it out for myself. This is all *your* fault.

For my mother, Blanche.

A life simply lived.

1 Introduction

Academic papers use citations to other papers in order to provide support for their arguments as well as to establish credit and provide historical context. The support is usually in the form of factual assertions that are developed in depth and defended in another publication [Weston, 2009]. The supporting papers also use citations leading to a network of citations upon which the argument rests. Support can be indirect (being several citations away) providing the evidence for assertions that is built on like the foundation of a house. As support flows across the network, ideas likewise flow and evolve in time. Progress can be impeded by a lack of support or new ideas available to the researcher, a situation that may arise if a particular research community is unaware of relevant work published elsewhere. It raises questions on the possibility of determining if two communities are working on the same topic and whether the similarity between two communities be measured.

First, a method for identifying communities is required. One method is by using the citations in academic papers to group together those papers that share citations. The authors are obviously aware of the papers they cite and feel it is necessary to include these sources in support of their argument. Second, similarity between communities can be expressed as the similarity between the content of the papers they publish. It seems obvious to postulate that, with the exception of jargon, discussion on the same topic should use some of the same words, although "false friends" and equivalence cases from differing fields (e.g. validity and reliability in psychology with accuracy and precision in physics) can be a confounding factor here. Thus, a method is required for measuring the similarity between documents.

The Complex Networks community has studied citation networks and also has algorithms for finding communities based on the links between individuals. The subject of Text Mining has methods for extracting measurable quantities from documents. In this thesis, I will use community finding algorithms on a citation network to identify communities that are aware of each other and use text mining techniques to calculate

the similarity between the documents in those communities and others as a proxy for the distance between communities.

In the absence of a widely accepted taxonomy of areas of knowledge, the following convention shall be used to refer to the scale of investigation. A topic is a single issue or small coherent collection of inquiries, a subject is made up of topics, a branch has many subjects and a field is made up of branches. This crude hierarchy breaks down somewhat when facets come into play, such as breaking the subject of Dark Matter into 4 groups which are collections of topics and yet referred to as topics themselves.

Table 1.1: A taxonomy of areas of study

field	branch	subject	topic
physics	astrophysics	Dark Matter	gravitational lensing
computer science	computer security	cryptography	cryptographic algorithm

1.1 Networks

In the last 17 years, complex networks approaches have been used in a wide variety of empirical studies, for example the collaboration networks of company directors, film actors and scientists, the power grid, the neural network of the worm *Caenorhabditis elegans*, epidemics and disease spreading networks, the World Wide Web and the network of sexual contacts [Newman et al., 2001, Watts and Strogatz, 1998, Ball et al., 1997, Liljeros et al., 2001]. While not the first, an important milestone for citation networks is Redner [1998], placing them in the context of complex networks, in order to benefit from the work of Girvan and Newman [2002] on the community structure of social and biological networks. Rosvall and Bergstrom [2008] is an excellent example of

using citation networks to map out the large-scale interdisciplinary structure of science. This thesis examines the other end of that spectrum, the multi-discipline study of a single subject.

1.2 Text Mining

"Big Data" is a boon to researchers who have developed tools for *data mining*, extracting patterns and learning rules buried in overwhelming swathes of numerical data [Witten et al., 2011, Callan, 2003]. Text mining uses the same concepts on documents, which are less structured, for tasks such as document classification of news articles, searching catalogues, information extraction and clustering [Manning et al., 2008, Weiss et al., 2010b]. Fortunately, large documents can be reduced to small representations for improved efficiency [Salton et al., 1975]. Information Retrieval has used clustering algorithms to group together similar documents to speed up searching and returning results from catalogues [Jardine and Rijsbergen, 1971].

1.3 The Relation to Knowledge

A broad community of researchers seek a better understanding of knowledge, how it is acquired and disseminated [Blackman and Benson, 2012, Bellotti, 2011, Börner et al., 2003, Rogers, 1962]. In contrast to the paradigm shift model introduced by Kuhn [1962] which proposes that change in science is a revolutionary process, the counter-argument by Toulmin [1972] that science is an *evolutionary* process is supported by recent studies on the social dynamics of science and how topics evolve [Jo et al., 2011, Choi et al., 2011, Sun et al., 2013]. There is considerable literature on Social Network Analysis and Communities of Practice [Groh and Fuchs, 2011, Friedkin, 1982, Yang et al., 2013, Lave, 1991, Wenger, 1999], but this thesis restricts itself to just the research outputs and not to the people involved. The act of citation is not personality-based, affected

by writing style, but is dependent on the textual content of the paper for its impact [Guerini et al., 2012]. By analysing the text, the flow of knowledge across the network can be quantified [Bhupatiraju et al., 2012]. Any overlap between fields can be traced through citation as an indicator of the cross-fertilisation present in interdisciplinary subjects [Chakraborty et al., 2014].

1.4 Research Questions

RQ 1: What is the relationship between the citation network and the textual content of a scientific research subject in terms of textual similarity within and between network communities found using the citations between scientific papers?

In their definition of Topical Community, Mei et al. [2008] proceed on the assumption that the structure of the network corresponds to the textual similarity of its members. Their objective is to improve the topic modelling as it applies to text summarisation, topic mapping and topical community finding, but this has not yet been shown. A positive answer to the research question provides support for their assumption, while a negative answer undermines the validity of their argument.

RQ 2: What is available in the textual content of a citation network that can help identify a scientific community?

Community-finding algorithms have had difficulty recovering the "ground-truth" of the communities through the metadata associated with them [Hric et al., 2014]. They have relied on the metadata tags applied to nodes in the networks, but this has proven unreliable. As the nodes in a citation network are documents, the similarities in the textual content of those documents are perhaps a better source of information about the true nature of the nodes than are externally chosen tags.

While the term scientific community could refer to research subjects or to the groups of scientists themselves, here, scientific community refers to the groups of papers authored by scientists and found on the citation network of a scientific research subject.

1.5 Summary

To undertake this study, it was decided to obtain a corpus of documents on a single subject for text processing and to create the citation network from those documents, their references and citations in order to examine and quantify the textual overlap of scientific communities. The subject of Dark Matter in the field of astrophysics was chosen as it has researchers from several different recognised disciplines studying the problem who deposit a significant proportion of their papers in arXiv, a large repository for physics and related fields described in Appendix A.2. The data collection process is described in Chapter 2 and a brief description of the Dark Matter problem is given in Appendix E.

As the approaches of complex networks and text mining are distinct, both disciplines have been given their own chapter. Chapter 3 starts with a brief history of complex networks, introduces the significance of communities, outlines the data collection method and presents preliminary results. Chapter 4 places text mining in the context of the many fields that manipulate text for aiding knowledge, such as Information Retrieval, Natural Language Processing and Artificial Intelligence. It then describes the process for reducing the documents to the Vector Space Model representation and presents preliminary results from k-means clustering, which was cross-validated. Chapter 5 combines the techniques of both to examine the overlap of the communities. The results of all three chapters are then discussed in Chapter 6 and suggestions made for future investigations in Chapter 7.

2 Data Acquisition and Curation

2.1 Motivation

To investigate the interrelations between communities in citation networks and the vocabulary used to express their textual content, a corpus of text was required and citation data for those documents obtained. Using background knowledge gained from my undergraduate degree in astrophysics and discussions with astronomers and astrophysicists, I chose the subject of Dark Matter to be the basis of the text corpus. The initial discussions highlighted three or four possible large topics within the subject that should provide sufficient differences in text for substantive analysis.

2.2 Text Corpus

One source of scholarly articles is the arXiv pre-print server for physics. Described in Appendix A.2, it is a facility for researchers to upload pre-prints, which are scholarly articles made available prior to publication. Many physicists use LaTeX to prepare their articles and arXiv allows them to upload the original sources of their articles, be that LaTeX PDF, PostScript, HTML, or Microsoft Word files.

There are disadvantages with using arXiv as a source. It only began in 1992, limiting the subjects available for study to relatively recent ones, before growing in popularity, meaning that coverage during the 1990s is less than what it is today. It is missing published articles because arXiv relies on researchers to voluntarily upload them to the service. Indeed, the service only exists because of its utility and has grown in popularity as researchers recognise its value in promoting work and gaining early access to current research.

2.2.1 Selection Criteria

The corpus was collected in 2009 from the arXiv search page. Two searches were performed using the search terms *"dark matter"* and *"MOND"*. Dark Matter is a popular term and researchers have incentive to use it to communicate to their audience how the paper relates to the subject. MOdified Newtonian Dynamics (MOND) (see Appendix E for a description), is a popular alternative to the Dark Matter hypothesis. Its purpose in the search process was to select papers related to Dark Matter in opposition which may not have been identified using the term *"dark matter"*. It was not expected that the search terms would necessarily recover all papers associated with the study of Dark Matter, only that a sample size sufficient for textual analysis would be obtained. In comparison with other sample sizes used for document clustering, while some recent studies have chosen to work with a corpus of more than 8000 documents [Huang et al., 2016, Tang et al., 2016], previous work has been satisfied with fewer than 2000 documents. Blei and Lafferty [2007] used 1452 documents and Steinbach et al. [2000] used an average of 1885 documents among 8 collections ranging in size from 690 to 3204. It was expected that gaps in the coverage of the Dark Matter topic would be filled through citation data, reducing the number of papers missing from the citation network analysis.

The strength of this curation technique is that given the search terms, any researcher can freely obtain the same papers used in the analysis. Without special access to the commercial property of several publishers, complete coverage of the topic could become very expensive, requiring subscriptions to journals beyond what the average institution may be willing to pay. It would also be impractical to reproduce the study. The original sources in a markup language such as LaTeX offer possibilities for further analysis of abstracts, equations, citations or other structural features found in the document.

Figure 2.1 shows the number of papers in the Dark Matter corpus submitted to arXiv each year. Its growth over time mirrors the rise in usage of the arXiv service as well as the increase in research in Dark Matter. In 2017, the search was repeated on

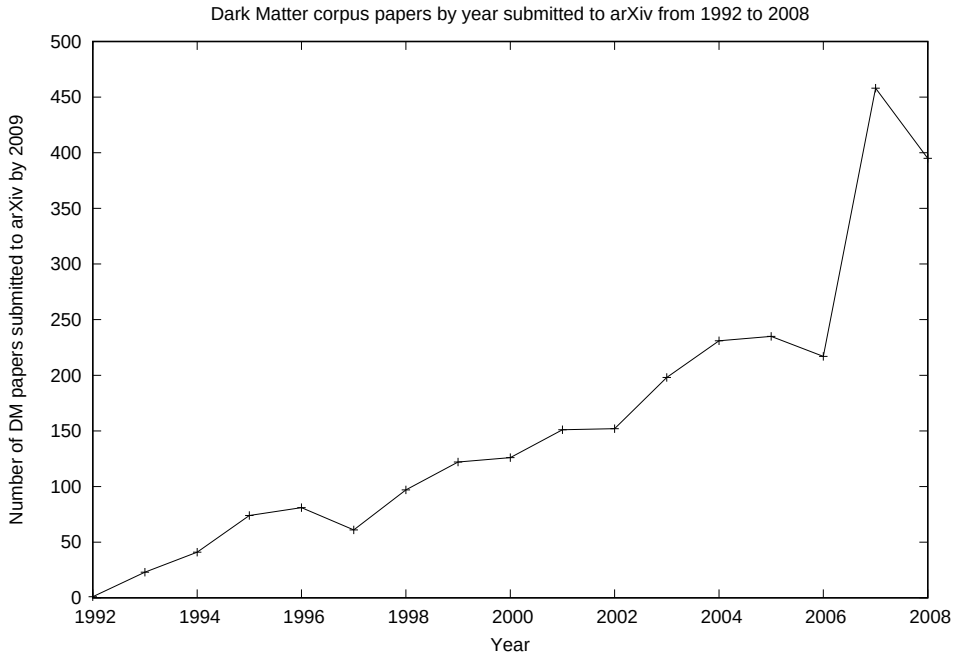

Figure 2.1: arXiv search results for Dark Matter in 2009 by year

both arXiv and Web of Science (WoS). Figure 2.2 shows the results of all 3 searches. Using the WoS as an estimate of all possible significant documents, arXiv's coverage of the Dark Matter topic in both 2009 and 2017 are plotted in Figure 2.3 as a ratio between the results found in arXiv and WoS by year. Currently arXiv holds more than half of the topic papers published after 1997.

Thomson Reuters' business model for WoS is to provide a comprehensive citation index for its subscribers. Their coverage encompasses over 50 000 books, 12 000 journals and 160 000 conference proceedings. Full coverage of a subject may never be obtained, but it has a commercial incentive to capture all available work that has impact and influence. It is for this reason that WoS has been used here as a proxy for full coverage.

To understand the size of the corpus in the context of the greater body of published work on Dark Matter, results for the search terms using the Web of Science [Falagas et al., 2008] were obtained, and are presented in Figure 2.2 broken down by

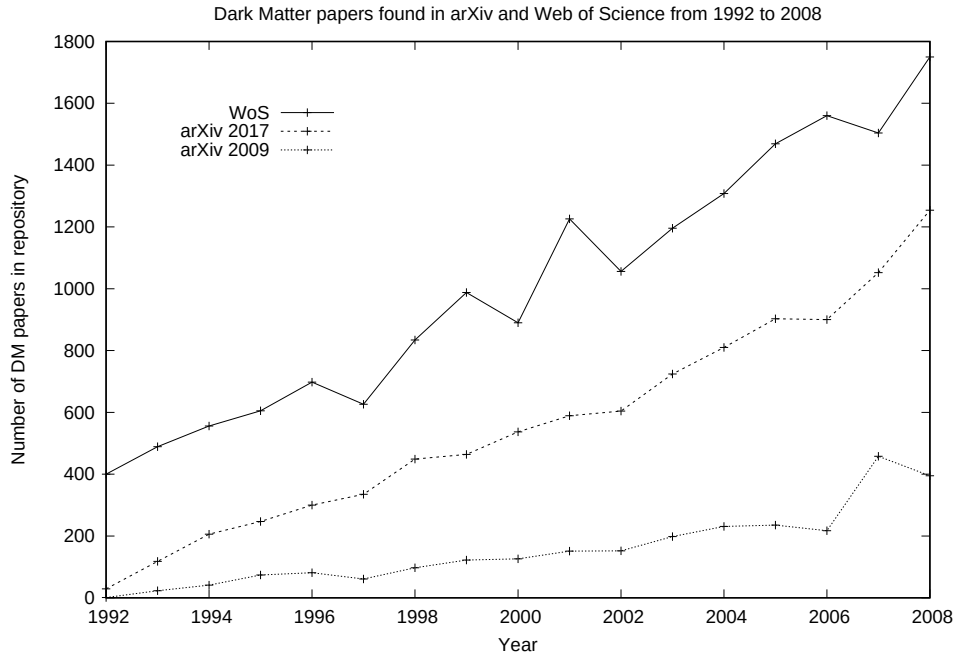

Figure 2.2: Web of Science search results for Dark Matter by year in comparison to arXiv search results in 2009 and 2017

year. It was found that for the years 1992–2008, Web of Science holds 16 508 papers corresponding to Dark Matter in the categories of `Astronomy & Astrophysics`, `Physics`, `Nuclear Physics` and `Instruments & Instrumentation`. The breakdown by year shows the sharp rise of Dark Matter papers lodged in arXiv against the background of increased publishing in Dark Matter from the inception of arXiv in 1991 to 1995 where it constitutes 10% of the WoS holdings and settles into a slow, steady increase, almost doubling its share of WoS in the following 13 years. These comparisons based on comparing calendar years will not be exact because publication is not necessarily the same year as submission to arXiv due to the time taken by the editorial process. The total submission rate in the years 1991–2009 is found on the arXiv website, arXiv.org [2010]

The first years of arXiv were dominated by submissions in High-Energy Physics (`hep`), a result of its origin at the Los Alamos National Laboratory, a major centre

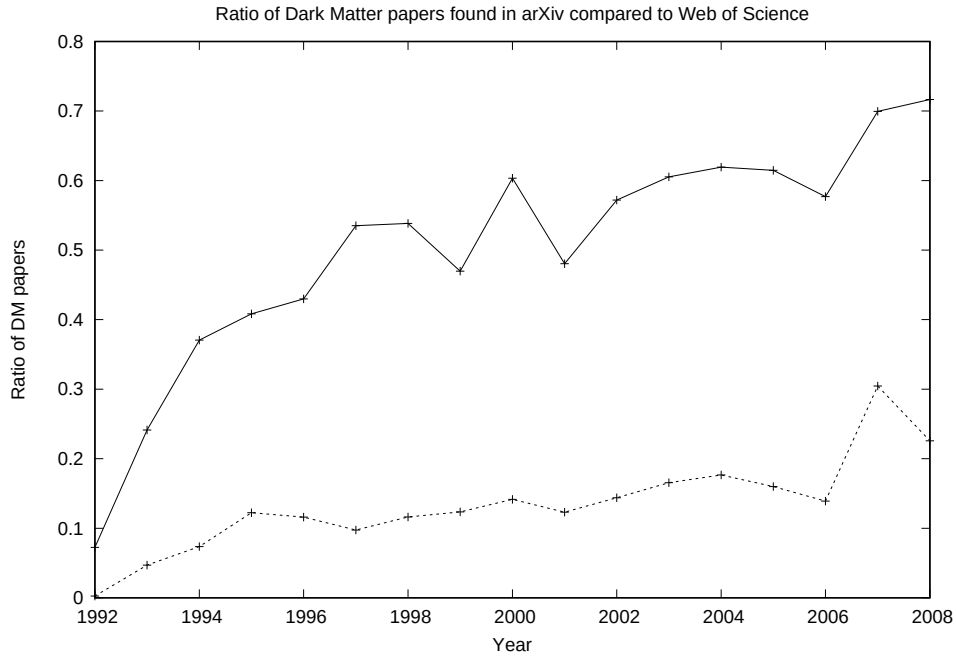

Figure 2.3: Estimated coverage of arXiv search results from 2009 and 2017 by year

in nuclear research in the United States. By 2005, `hep` was no longer receiving the majority of submissions as the astrophysics and condensed matter research communities took advantage of this freely-available, government-supported service. The yearly breakdown of the Dark Matter corpus in Figure 2.1 reflect the steady adoption of arXiv by the astrophysics community.

The search results found 2671 which were downloaded from arXiv. As noted in Table 2.1, 8 papers were removed because they were not involved in Dark Matter and 2 papers were removed from the corpus because their content consists only of the notice that the submission has been removed from arXiv. The remaining 2661 papers selected from arXiv constitute what will be referred to here as the Dark Matter corpus. Appendix F.2 contains a list of arXiv ids from 1992 to 2008 that make up the Dark Matter corpus.

A researcher today will get a much larger sample of papers from the same search

terms in arXiv. While it is possible that the search technology has drastically changed in 8 years, more plausible explanations are that a push towards making more research *open access* has led publishers to allow pre-prints to be uploaded or that as arXiv becomes an integral part of the research culture, researchers see the value of uploading older papers so that they, too, are freely accessible.

2.3 Citation Network

The citation network was built around the citations to and references from the 2661 papers of the Dark Matter corpus. It was found that the Astrophysical Data Service (ADS) (described in Appendix A.1) had excellent excellent coverage of the citation data for this subject and had an API for fetching the data. The arXiv IDs identifying the papers of the corpus were translated into the "bibcode" identifiers used by ADS. The list of bibcodes was placed in a queue stored in a database. A Perl script removed a bibcode from the queue, made a query to fetch the paper's metadata (title, authors, keywords, journal, date published) and then made two more requests to fetch the list of references the paper has and the list of papers identified as citing this one at that time. The queries using the ADS API were made every 10 minutes to avoid overloading a stretched resource for the astrophysical research community. Upon failure of a query, the script would back off, and double its wait between requests. Whilst running continuously unattended, this script needed to be able to handle service disruptions, network failures and power outages. It would resume its regular schedule on a successful query. The backing-off behaviour had the benefit of reducing the network traffic and log entries during disruptions, while minimising the time between service being restored and the data fetch resuming. The references and the citations were stored in separate database tables in order to ensure that the direction of the citation relationship was preserved. At the end of the data collection, the two database tables were consolidated into one table (duplicate entries were removed) holding 7 606 982 links from which graph representations could be produced for analysis.

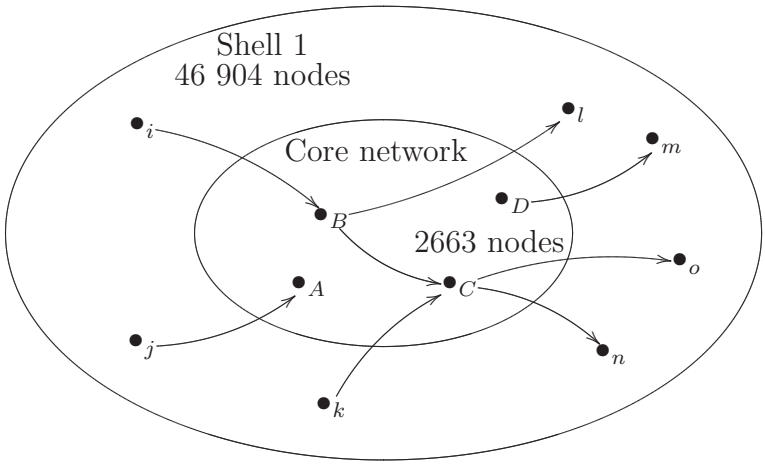

Figure 2.4: Shell 1 – Citations to and References from the Dark Matter core

The initial citation network was constructed after 24 371 web queries were made in a little over 15 months from June 2011 to October 2012. After all citation and reference bibcodes to the Dark Matter corpus had been collected, an initial analysis showed a discrepancy in the degree distribution compared to other citation networks. A decision was made to repeat the process for the newly added bibcodes to ascertain the nature of this discrepancy. The second run was completed in just under 2 years, making 139 321 web queries from October 2012 to September 2014.

The citation network can be subdivided into 3 regions according to their citation relation to the Dark Matter corpus: the *core network* consists of all nodes representing the Dark Matter corpus. The *first shell* surrounding the core network is made up of the nodes that are either references from or citations to nodes in the core network, and the *second shell* refers to the remaining nodes that are references from or citations to nodes in the first shell.

In Figure 2.4, nodes A, B, C and D inside the inner ellipse represent papers in the core network. An internal citation is shown between papers B and C in the Dark

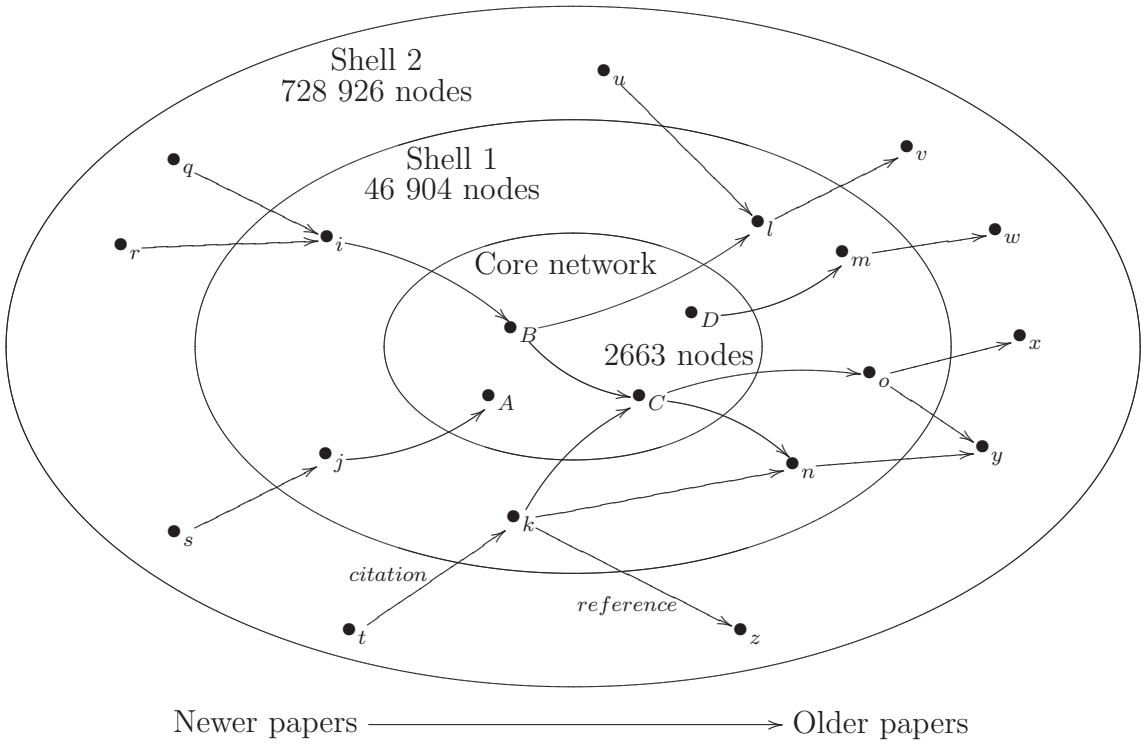

Figure 2.5: Shell 2 – Citations to and References from nodes in Shell 1. Curved arrows are citations found in the first pass, straight arrows are citations found in the second pass (notice k - n)

Matter corpus. External citations (nodes i, j and k) and references (nodes l, m, n and o) are made to papers that will become the first shell of the network. In Figure 2.5, citations internal to the first shell are made (nodes k to n), external citations (nodes q, r, s, t and u) and references (nodes v, w, x, y and z) are made to papers now in the second shell. There are no citations from the second shell to the core network, otherwise they would have been discovered during the first phase of the data fetch and hence been allocated to the first shell. As only minimal data regarding the second shell has been retrieved, via their links to the first shell, they play the role of supporting actors, or the scaffolding on which the citation network is built.

2.4 Document Processing

A compressed tar archive of the papers (colloquially referred to as a "tarball") consisting of the Dark Matter corpus was provided on request by the arXiv administrators. The tarball was unpacked into a file structure where all the source files the author used were contained within a separate directory for each arXiv submission. The submission can be in several formats. Table 2.1 presents a breakdown of the file types (LaTeX, PDF, PostScript, HTML and Microsoft Word) found in the corpus. The files in the directory were tested with both the Unix `file` command and by examining the file extensions to determine how they should be processed to extract the text.

Because LaTeX is an extremely popular markup language in science and academia and the language emphasises a logical rather than visual layout, science articles tend to have an *abstract* section which can be located and accessed separately with tools such as LaTeXML (described in Appendix B). This package of Perl modules was used to extract all the passages of text identified by the XML tags `<p>` or `<text>` to be processed as the *textbody* section, `<abstract>` to be processed as the *abstract* section and `<keywords>` to be processed as the *keywords* section. This effectively removes mathematical notations from the processing stream. `pdftotext` converts all text in a PDF file to text, meaning that any equations will be included in the text. Filtering

papers	*file type*	
2 510	`.tex`	LaTeX
73	`.pdf`	PDF
73	`.ps`	PostScript
3	`.html`	HTML
2659		*subtotal of papers processed*
2	`.docx`	Microsoft Word format
2661		*subtotal of papers curated*
2	removed	(1 plagiarism/1 author removal)
8	`math` category	removed because of author's name
2671		*total files received from arXiv*

Table 2.1: File types of Papers in the Dark Matter corpus received from arXiv. 2671 papers received, 2659 papers processed.

out single letters from processing reduces the noise from this source. PostScript files are first converted to PDF using `ps2pdf` and then processed as PDF files. HTML is rendered to text using the `parse` method of Perl's `HTML::Strip` module. It removes all HTML markup and returns plain text. As programs such as `latex2html` render mathematical markup as images which are removed by stripping, the noise introduced by equations should be lessened.

To process the text into VSM files, the text is split into tokens on the non-word character boundaries (non-letter and non-digit), meaning that any contiguous string of alpha-numeric characters is considered a "token". This strips out punctuation. To remove the extraneous possessives created by **'s** being converted into a single **s**, single letters were removed. Tokens of less than 4 characters that include a lowercase letter were removed. The number "3" is removed because of its short length. The word "three" is retained. A quantity such as $1.6 \times 10^{29} kg$ is rendered in LaTeX's math mode which LaTeXML encapsulates in a `<Math>` tag. These are excluded from text processing. The rationale for the decision to remove numerical values is that they only contain information when presented in context, something that is removed by the VSM model. In order not to remove acronyms, tokens consisting entirely of uppercase characters

were retained. Stopwords (described in Section 4.3.2 and tabulated in Tables F.1 and F.2) were removed from the list. Since the choice of stopwords is not critical to the results, Perl's standard stopword list, `Lingua::EN::StopWords`, was considered an acceptable starting point. The list was augmented by 41 terms from the top 200 most common words in the Dark Matter corpus. They were selected on the basis that their inclusion adds little to the meaning, many being abbreviations such as *et. al., fig.* for figure and *eq.* for equation. The rest are common to scientific articles, such as the use of the past passive tense. Multiword tokens were not considered, a choice which may increase the noise slightly. Normalisation was not undertaken for the same reason of choosing simplicity over the small improvement provided by perfection. Reference lists at the end of the documents are encapsulated by a `<bibliography>` tag and are not included in the text processing. The remaining tokens were stemmed using Porter's algorithm (described in Section 4.3.4). As the tokens are extracted, the number of occurrences of each token in the file are counted and are collected into a frequency file, labelled as textbody, abstract or keyword, for future rendering as Vector Space Model (VSM) representations. For faster loading and reduced computer memory consumption during clustering, the frequency files were saved as Berkeley DB files, a database file format that Perl can treat as an in-memory data structure.

A small number of files were packaged in a non-standard fashion, such as confounding file extensions (e.g. `.bak` and `.ltx`), or extra front-matter added by email systems that mislead the `file` command into classifying the document as ASCII text instead of its correct file type (LaTeX, PDF or PostScript). They were identified manually and the script consulted a lookup table to determine which processing was required.

2.5 Summary

These steps are also partially described in the following chapters as the theory is introduced to illustrate the purpose for choosing to collect and process the data as outlined in this chapter.

3 Citation Networks

3.1 Introduction to Networks

The popular awareness of the small world phenomenon, "Six Degrees of Separation" is believed by many to have originated from Stanley Milgram's experiments on the small world problem in the late 1960s [Milgram, 1967, Travers and Milgram, 1969]. While Milgram found the average number of intermediates between two strangers to be 5.2, he did not use the phrase "Six Degrees of Separation". That entered the public consciousness as the title for John Guare's extremely successful Broadway play, later made into a movie. The provenance of the phrase is light-heartedly investigated by Albert-László Barabási in his 2002 book, *Linked*, [Barabási, 2002] where he makes a compelling case that the concept is, like himself, Hungarian, tracing it to a 1929 short story by celebrated Hungarian author, Frigyes Karinthy, called "Láncszemek" or "Chain-Links" where the protagonist proposed the likelihood of linking himself to anyone in the world in 5 acquaintances. He has even included an English translation of it in a 2006 collection of significant research papers on networks, co-edited by Mark Newman and Duncan Watts, as a historical footnote [Newman et al., 2006]. Both *Linked* and *The Structure and Dynamics of Networks* provide an excellent historical overview of the development of complex networks and small worlds.

A short summary of the main historical events often found in the introduction section of research papers on the topic frequently begin with Euler inventing graph theory in 1736 to solve the Königsberg Bridge Problem, a mathematical puzzle popular with the citizens of that Prussian city (now called Kaliningrad in Russia [Newman et al., 2006, p. 1]), to discover a continuous path crossing each of seven bridges linking the two sides of the river and two islands only once. Euler proved that, since a node in a graph with an odd number of links must be a starting or ending point of a continuous path, no such path across the bridges of Königsberg existed. This is the recognised beginning of Graph Theory and the work that followed dealt with regular graphs, like

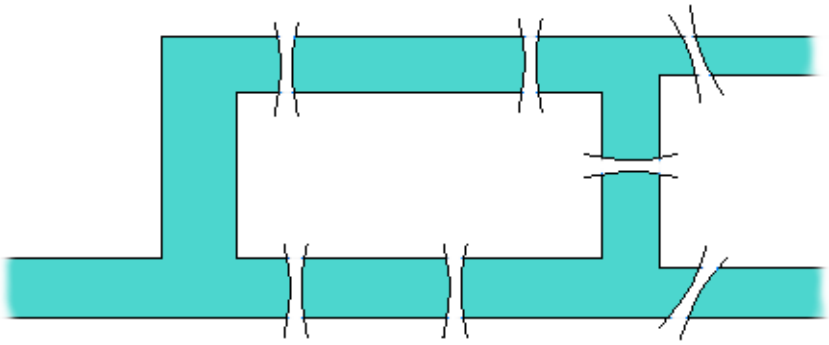

Figure 3.1: A diagram of the Königsberg Bridge Problem

lattices and other repeating patterns.

Investigation into Random Graphs grew rapidly in the 1950s with the work by Paul Erdős and Alfréd Rényi hailed as the seminal work. Although re-discovering results obtained a decade earlier when Solomonoff and Rapoport examined what happened to a graph whose nodes were connected at random in trying to model a number of biological scenarios, the 8 papers by Erdős and Rényi from 1959 to 1968 set out many of the important features of random graphs. The main feature of the Erdős-Rényi graph is its small world property, that despite a large number of nodes in the graph, the number of links separating any two nodes is small on average.

In the same years, sociologists had also been trying to understand the world through the social networks we create via friendships and acquaintances. Inspired by Ithiel de Sola Pool and Harrison White, Milgram's famous experiment sent a request to individuals randomly selected from the phone-book of Omaha, Nebraska to pass a letter onward to a named stockbroker over 2000 km away in Massachusetts only via people they knew on a first-name basis. The mean number of 5.2 steps between individuals was vastly lower than his colleague's estimate of 100 steps and showed the real world importance of the study of random graphs.

It was not called networking in the early 1970s when Mark Granovetter set out to uncover the sociology of finding employment. Rejected for publication the first time, the surprising result outlined in *The Strength of Weak Ties* was that people were more

successful finding jobs through acquaintances than through close friends [Granovetter, 1973]. This early paper, grounded in empirical data, emphasised the significance of community structure in the composition of the network, something absent in the random graphs of Erdős and Rényi. The random nature of the links does not promote clustering of the nodes.

Two decades later, mathematicians Duncan Watts and Steve Strogatz began looking at the small-world problem including the studies by Milgram and, crucially, Granovetter. Starting with a highly clustered, yet large-world network, they found that, by randomly re-wiring less than 1% of the links, the distance between nodes dropped dramatically. They published their findings in 1998, showing how a small world appears out of mostly ordered networks using comparisons of the actors network, the electrical power grid and the neural map of *C. elegans* to similar sized random graphs [Watts and Strogatz, 1998].

A year later, Albert-László Barabási and Réka Albert published results from a study of the structure of the World Wide Web, a network defined by the hyperlinks between web-pages [Barabási and Albert, 1999]. They showed that this network as well as the *C. elegans* neural map provided by Watts contained hubs (nodes with many more links than average), a feature not consistent with either Erdős-Rényi and Watts and Strogatz models. The number of links attached to a node is called its "degree". In plotting the distribution of degrees, they found it followed a power law relation and suggested that power laws might be a generic feature of complex networks. They also noted that the WWW is a growing network and proposed a model for the observed distribution called "preferential attachment".

It has since gone on to be a wildly popular technique and has been applied to: physical infrastructure (the topology of the Internet), biology (protein-protein metabolic networks and epidemic spreading), sociology (social network analysis and criminology), finance (the World Trade Network) and transportation (the airline network between airports).

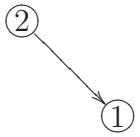

Figure 3.2: The author's paper[2] references the original work[1].

Presented here is a simple yet complete explanation of the graph theoretic definition of a citation network. In mathematics, a graph is formed by a set of "vertices" which can be connected to each other by a collection of "edges". This graph has no geometrical significance. Its only purpose is to show how entities are connected. Unfortunately, the terminology is not standardised and commonly vertex is replaced with "node", edge is replaced with "link" and "network" is used interchangeably with graph. For the remainder of this work, the the terms **nodes** and **links** will be used.

With respect to the process of authoring an academic paper, a responsible author attributes credit to previous work and provides evidence for arguments presented in a paper by adding a "citation" to the original work in the References section at the end of their paper, such as seen in this thesis.

The graph created by that citation has two nodes, representing the author's paper and the original work connected by one link, with the citation going *from* the author's paper *to* the original work, as in Figure 3.2. There is no citation in the original work to the author's paper because it did not exist at the time of the original work and having been published, it cannot be changed. This is called a "directed" graph and each node now has two types of degree. The in-degree is the number of links directed to the node and the out-degree is the number of links emerging from the node.

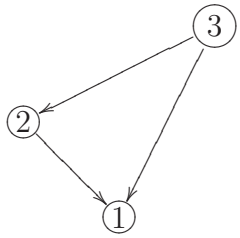

Figure 3.3: A third paper[3] references the author's paper[2] and the original work[1].

In graph theory, a "path" is a set of nodes traversed via their links. Starting at a node on a graph, if it is possible to follow one of its links to the next node (obeying the link direction if that is defined) and continue following links until arriving back at the starting node, then that path is called a "cycle". Citation networks should never have such a cycle due to the unchanging arrow of time and hence are called "acyclic".

Should another paper be written discussing results from both of the papers discussed above, it can in theory reference both papers. Now the simple graph appears as in Figure 3.3 where node 1 has in-degree of 2, out-degree of 0 and node 2 has in-degree of 1, out-degree of 1. Time has passed and the network has gotten bigger. This is a "growing network" as new papers that cite earlier work are constantly being written, so more nodes and links are being added all the time.

A smaller section of a graph is called a "sub-graph" and if all the nodes in that sub-graph are connected to every other node in the sub-graph, it is called "complete". The smallest example of this is the three papers of Figure 3.3, which are all linked together by citations. This complete sub-graph is called a "clique", just one of the many terms for this structure arising from the work on social networks and is used in measuring the amount of clustering in a network.

Links, once added, are static and remain fixed in place in contrast to hyperlinks on the WWW which can be changed or removed at any time according to their author's whims. This description overlooks the case where a pre-print is cited (the author has had an advanced view of a yet-to-be-published paper) and the link is directed forward in time and ignores e-publications which could conceivably be edited after its publication.

In summary, a citation network is a *directed acyclic graph* (DAG) of academic papers linked by their citations which is a *growing network* and has *static links* that are almost always directed at older papers.

While Erdős, Rényi and Milgram were working on small worlds, historian of science, Derek de Solla Price, took advantage of the machine-handled citation data in Eugene Garfield's 1961 Index to extract the power law distribution from the citation network. He later proposed that the power law could be generated from a "cumulative advantage" distribution (called preferential attachment by Barabási and Albert) based on work by Simon [Price, 1976, Simon, 1955]. The significance of these results was not widely noted outside the Information Science community until after Barabási and Albert's work on the WWW produced similar features.

This is not to say that physicists were ignoring citation patterns and publication trends. Jan Vlachý, editor of the Czechoslovakian Journal of Physics, who uses citations

to look for paradigm changes had cited Price many times and was awarded the Derek de Solla Price Medal in 1989 for outstanding contributions to the field of quantitative studies of science. Conversely, Helmut Abt, astronomer and managing editor for the Astrophysical Journal, reported regularly on such themes as the exponential growth of the literature and the "half-life" of a paper, but never cited Price [Abt, 1981, 1984, 1987, 1992, 1996, 1998b, 2006]. It is not known if either of them examined the citation distribution.

Independently in 1998, Steve Redner re-discovered the power law distribution using data from the Institute for Scientific Information (ISI) and *Physical Review D* [Redner, 1998]. It is not unusual for researchers to arrive at the same or similar conclusions unaware of previous or concurrent work in the same area, as was the case with Erdős and Rényi and many others [Simkin and Roychowdhury, 2011]. Perhaps it is the timing of his re-discovery that is so opportune, coming only a year before Barabási and Albert used it in their mounting evidence for the importance of cumulative advantage in complex networks.

3.1.1 Common Features of Complex Networks

The three significant characteristics of a complex network are the "small world" property, clustering and a fat tail. As in the random graphs of Erdős and Rényi, it should take few steps between any two nodes compared to the size of the network. The increase in the mean path length should be proportional to the increase of the *logarithm* of the network size (or less). Many networks in the real world display more clustering together of nodes than would be expected from random connections. This is the beginning of the complexity of the network. The fat tail signals the presence of hubs (nodes with a much higher number of connections) and is associated with power law relations. These relations are also known as "scale-free" because similar structures appear at all levels. There is no characteristic "scale" for the network as it looks the same whether zoomed-in or zoomed-out. It is this that gives the network its complex topology. Plotting the proportion of nodes with a given degree against the degree on

a double-log scale usually leads to a straight line. The slope of this line yields the exponent of a power law from which conclusions can be drawn about the nature of the connection process.

Complex networks can be differentiated on attributes dependent on how links are made. Is there a direction associated with the link, are some links stronger than others, are new links being made or existing links being removed? These questions help to find comparable networks, regardless of domain.

3.1.2 Clustering

One of the characteristic features of complex network is the tendency for the nodes to cluster together in terms of an increased sharing of connections between nodes. To measure the amount of localised clustering in a network, the clustering coefficient, C_i, for a node i is defined as the ratio of the neighbours that are linked, e, to the maximum number of possible interconnections [Caldarelli and Vespignani, 2007]. To derive the maximum number of connections between k nodes, each node can be connected to the remaining $k-1$ nodes. Doing this k times gives us $k(k-1)$, but this product counts each connection twice, once from i to j and a second time in the reverse direction from j to i. In this way, we see that the maximum number of links is $k(k-1)/2$ which yields the relation

$$C_i = \frac{2e_i}{k_i(k_i-1)} \tag{3.1}$$

By definition, this statistic must always lie somewhere in the range $[0,1]$. Although, according to Dorogovtsev this definition is only well defined for undirected networks, it can also be applied to directed networks by disregarding their directedness [Dorogovtsev and Mendes, 2003].

With Erdős and Rényi considering that graphs were stochastic objects rather than purely deterministic, the graphs were treated as probability distributions [Newman et al., 2006, p. 4]. Certainly now, most characteristics for a network are described in terms of distribution functions.

Plotting the probability distribution for the clustering coefficients on the Dark Matter citation network in Figure 3.4 and described in Section 3.2.3 shows a curve that peaks around 0.14 for the core network, dropping to 0.10 for the larger full network and slowly decays as C increases. While at first glance it looks like a Poisson distribution, it has been difficult to fit such a curve to the data, nor does a Gaussian accurately describe it. An exponential subtracted from a quadratic came closest to replicating the curve but could not replicate the peak's height while matching the fat tail. While the distribution in the range $[0.2, 0.8]$ is adequately described by a power law of the form $P(C) \propto C^{-\gamma}$, most researchers only calculate the average value of C, written as $\overline{C}$, to describe the tendency for clustering over all nodes in the network.

One exception has been Colomer-de-Simón and Boguñá. In their modelling, they plotted $\overline{C}$ for artificially wired networks with different characteristics. Those plots show that, for lower average values of C, the value of C remains constant as degree k increases, until it encounters a power-law decay that was common across all their constructed networks [Colomer-de Simón and Boguñá, 2014]. As the Dark Matter citation network exhibits higher clustering at low degrees than their simulations, the power law behaviour of the plot of C vs k in Figure 3.6 extends across almost the whole plot with only a barely perceptible flattening at the lowest degrees. This shows that papers with fewer citations have a stronger tendency to share their citations, whereas highly cited papers are less likely to be cited together. There is a small but noticeable difference in the clustering coefficient for the network at different scales. The smaller core network generally has higher values of $\overline{C}$ and the full network at the largest scale (core plus shells 1 and 2) has the lower values, until the data blurs together around $k \approx 70$.

A deeper analysis of clustering on directed networks was done by Fagiolo [2007] in order to better understand the effect of weights on the clustering coefficient. (Weighted networks are not considered in this thesis.) He goes on to categorise the different types of triangle possible between nodes. He also plots C vs k for his World Trade Network which is highly clustered ($C \approx 0.8$) in comparison to the Dark Matter network.

3.1.3 Formal Definition of Complex Networks

As mentioned in Section 3.1.1, complex networks exhibit the **small-world property** in common with the random graphs of Erdős and Rényi. The mean path length of a complex network should be comparable to the mean path length of an equivalent random graph containing the same number of nodes, such that

$$\frac{\langle l \rangle}{\langle l_r \rangle} \sim 1$$

where the mean path length of a random graph is given by

$$\langle l_r \rangle = \frac{ln(N)}{ln(\langle k \rangle)}$$

I shall consider a network with $\frac{\langle l \rangle}{\langle l_r \rangle} < 2$ to have the requisite small-world property.

A complex network must demonstrate significant **clustering**, such that the clustering coefficient of the network is greater than the clustering co-efficient of an equivalent random graph [Watts and Strogatz, 1998]. The ratio of the two co-efficients are much greater than unity.

$$\frac{C}{C_r} \gg 1$$

where the clustering co-efficient for a random graph is given by

$$C_r = \frac{\langle k \rangle}{N-1}$$

For the purposes of this thesis, a clustering ratio of $\frac{C}{C_r} > 10$ shall indicate significant clustering to satisfy the second requirement for a complex network.

The third characteristic of a complex network is a **scale-free degree distribution**, commonly fitted with a power law, $P(k) \propto k^{-\gamma}$, where typical values for γ lie close to the values of 2 or 3. To establish what is not a complex network, the ration of the degree distribution at different scales is used.

$$P(k) \propto k^{-\gamma}$$

$$P(k) = Ak^{-\gamma}$$

$$P(10k) = A(10k)^{-\gamma} = 10^{-\gamma}Ak^{-\gamma}$$

$$\frac{P(10k)}{P(k)} = \frac{10^{-\gamma}Ak^{-\gamma}}{Ak^{-\gamma}} = 10^{-\gamma}$$

For the purposes of this thesis, I constrain the values for the power law exponent, γ, to be between 1 and 4, such that

$$10^{-4} < \frac{P(10k)}{P(k)} < 10^{-1}$$

This simplifies the discussion of which networks possess a fat-tailed distribution by calculating the ratio of the degree distributions at $k = 100$ and $k = 10$ and excluding distributions with ratios outside the range indicated above.

3.2 Acquiring the Data

To study the community structure of a citation network, a topic undergoing intensive study was chosen such that it should have at least 3 distinct communities of academics researching the problem. The study of Dark Matter has increased steadily since it was observed in 1969 by Vera Rubin and Kent Ford [Rubin and Ford, 1970]. It was believed that Dark Matter researchers could be classified into 3 or more communities as discussed in Section 3.3.1. A description of the research problem can be found in Appendix E.

It was decided to use the open access repository for physics, arXiv (See Appendix A.2) in order to be independent from the ISI, which is a substantial source of citation data, but also a commercial interest. In using the arXiv open access repository, it was assured that the text of the publications could be obtained without concern about restrictions from commercial publishers. It also places this study in a position to note discrepancies between the two data sources, should any arise.

3.2.1 Selection Criteria

The selection criterion for the papers was the search results returned by arXiv for the keyword searches on dark matter and MOND. The list was reduced when it was noticed that some of the MOND results were about lunar astronomy, rather than modified Newtonian dynamics and were mixed in because "mond" is also the German word for "moon".

The administrators of arXiv were contacted for permission to download the 2671 papers corresponding to the search results. They kindly packaged them together as one large download, rather than overloading their servers with the many requests to their web servers. On unpacking the download, the few papers in the `math` category were inspected and found that the author's surname was Mond and has no academic connection to Dark Matter. These papers were then excluded from the analysis of the network.

While searching for stop-word lists specifically in astronomy, it was found that NASA's ADS (see Appendix A.1) collects and makes available both reference and citation data [Accomazzi et al., 2000] and has a Perl module [Allen, 2003] for downloading data from the ADS. It was decided to seek permission to source citation data from the ADS in order to expand the network and be able to more accurately assess the in-degree of each paper, something that would not have been available solely through the reference sections of the papers in arXiv.

After negotiating with the ADS manager, the process of downloading reference and citation data from the ADS mirror in Nottingham was allowed at a rate of 1 request every 10 minutes so that the server was still able to respond to queries from other users of the service. This was permitted on the condition that the citation network itself was not to be published, a restriction imposed on the ADS by publishers who had provided them with their own proprietary data.

3.2.2 Data Collection

To collect the data over many months, a script was written that was resilient to restarting and stored the downloaded data in a MySQL database described in Appendix C. In using the `Astro::ADS` module, two bugs were found in the code. The first was due to the incorrect handling of publication identifiers, called "bibcodes", being sent over HTTP. The other bug was simply that the last reference returned by the ADS was being discarded, meaning that some citations would be missing.

Both bugs were fixed and patches sent to the module author in order that he could update the publicly available code repository known as CPAN (Comprehensive Perl Archive Network) from which `Astro::ADS` is distributed. This process concluded with me being given maintenance responsibility for the `Astro::ADS` module on CPAN. The patches are included in Appendix C.

After all references and citations to the core Dark Matter papers were collected, the degree distribution was plotted as described in Section 3.2.3 and the exponent found to be less than 2. This was a large departure from the values found by Redner [1998] and others for citation networks. A request was made to the ADS to download more citation data to determine if the low value was an effect of the network size or the data source.

When the data collection was finally finished, it was found that 23 of the core papers still did not have references. Upon investigation, it was discovered that 36 papers in the collection were missing citation data in the ADS. The reference section for each of the 36 papers was examined and references were located using the ADS search engine. The experience of manually locating references was informative of the difficulties in producing a clean and complete citation network. Many of these papers had different referencing styles that makes automatic extraction problematic. Errors and incomplete information in the paper at times prevented a secure identification of the correct bibcode. For example in bibcode `2006gr.qc.....6058D`, one reference to a 1917 paper by Einstein included all the necessary information with the exception of the page number or title. Usually that would be sufficient to identify the reference,

but in this instance, Einstein had authored 2 articles in the same issue of the same journal. Some references to books or literature outside of the normal realm of the ADS did not have existing bibcodes. While there exists a form to submit new bibcodes for these items, it was felt that this would not improve the citation network as most of the references to these items would be missing from the data already collected. There is also a form to submit the bibcodes of missing references. Of the 691 missing references, 561 bibcodes were located and submitted to the ADS, thereby improving the bibliometric coverage of the ADS for future users as well as the completeness of the Dark Matter network.

3.2.3 Methods

To facilitate the calculations, a Perl `Graph` object was built from the collected data in the MySQL tables from which duplicated entries were removed. The `Graph` object has methods for reporting the in-degree and out-degree for each node. It can also implement Dijkstra's algorithm for single-source shortest paths, but it was not used.

To find $P(k_i)$, the probability of a node having in-degree k_i, the number of nodes with in-degree k_i was counted and the distribution normalised by dividing all counts by the total number of nodes. $P(k_i)$ v k_i was plotted on a log-log plot in Figure 3.8 and a power law of the form, $k_i^{-\gamma}$ was fitted using `gnuplot` to the middle range of data where the points lie roughly on a line. From this fit, the power law exponent, γ_i, was estimated as described in Clauset et al. [2007].

The clustering distribution (Figure 3.4) was found by calculating the clustering value of each node using equation 3.1 and repeating the same procedure as with the degree distribution, $P(k_i)$, to find $P(C)$.

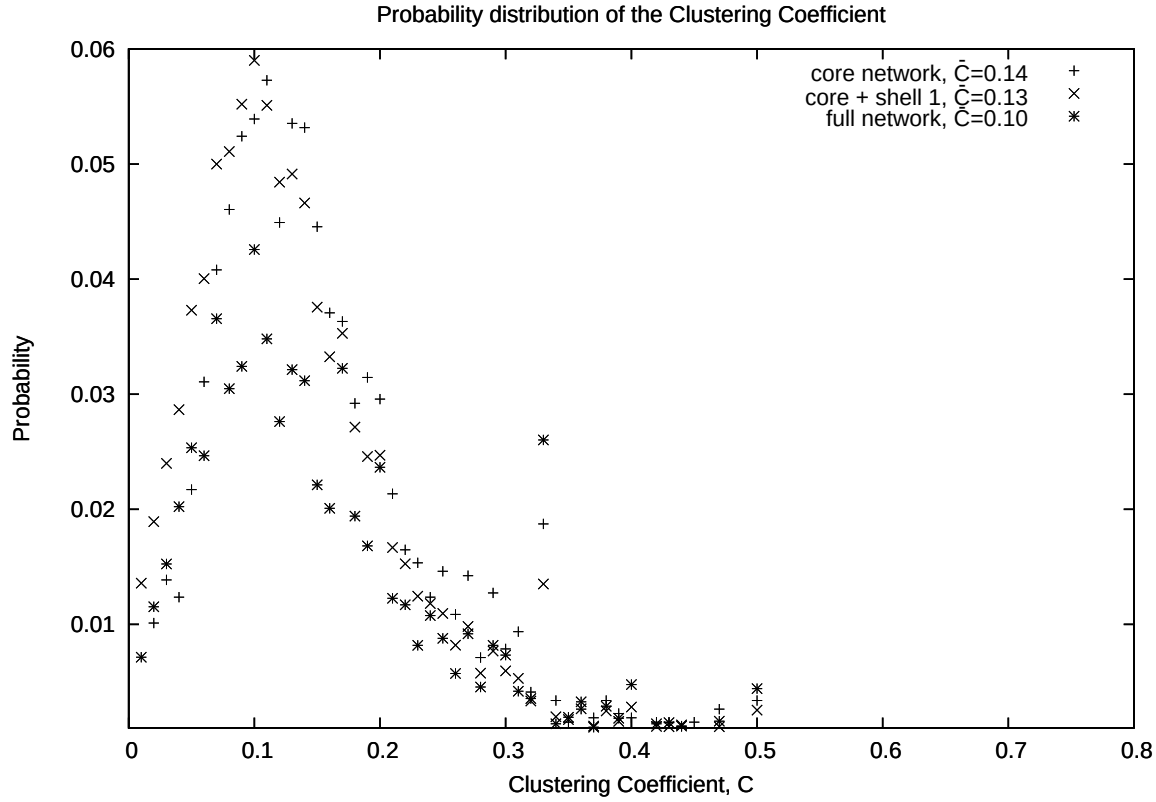

Figure 3.4: Distribution of clustering coefficient for different sizes of the Dark Matter citation network. The core network, core + shell 1 and full network listed in the key along with their average clustering coefficient, $\bar{C}$, are described in Figure 2.5.

3.3 Community

It seems plausible that the term *communities* used in network analysis would originate from the sociological studies of interpersonal relationships of neighbourhood residents of towns and cities in the 1960s and 1970s [Scott, 2000]. Researchers investigating the social organisation of selected populations found structures within their networks at different scales. Local communities found at a mesoscopic level are groupings that have a higher density of connections within the group than they do to external members. This rather vague generalisation is perhaps due to the natural language usage

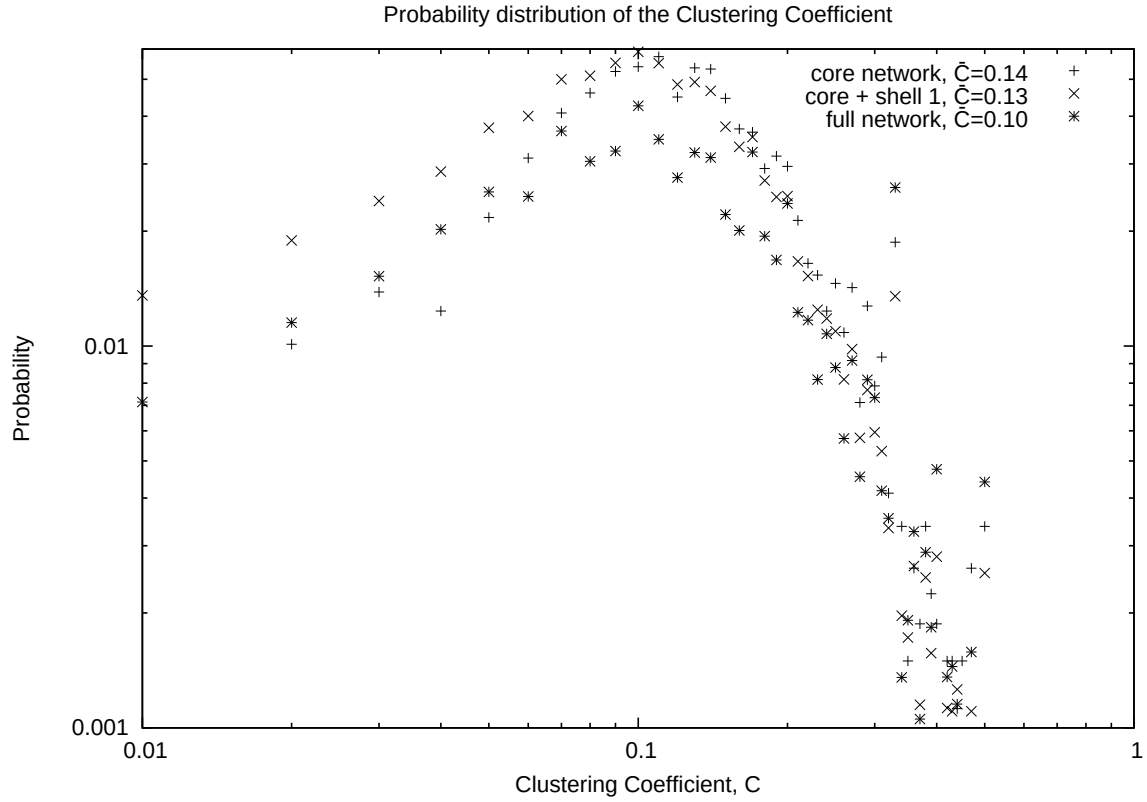

Figure 3.5: Log-scale distribution of clustering coefficient for different sizes of the Dark Matter citation network. The core network, core + shell 1 and full network listed in the key along with their average clustering coefficient, $\bar{C}$, are described in Figure 2.5.

of the term. Indeed, Danon et al. [2007] state that a consensus on a clear definition for *community* has not been reached in spite of much study in the area. Adding to the difficulty, communities may divide into smaller sub-communities. By extension, some networks may exhibit a hierarchy of communities connected by the measure of betweenness centrality of links connecting the nodes of the graph.

A smaller component of a community which in contrast has a simple yet rigorous definition, the *clique* is a complete subgraph, a subset of nodes in the graph that are fully connected with each other. The most basic clique is the triangle, as it is three nodes connected all connected together, as demonstrated in the example in Figure 3.3.

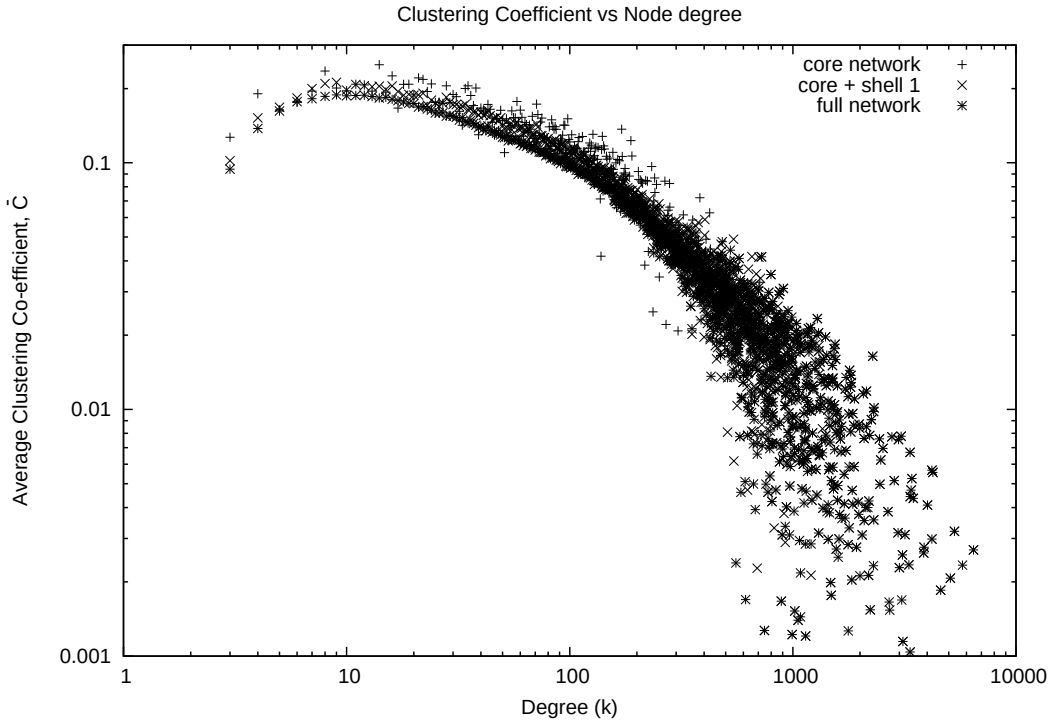

Figure 3.6: Clustering coefficient as a function of node degree, k for the network at three different scales, using 2659 nodes in the core network, 49 565 nodes in core + shell 1 and 778 491 nodes in the full network.

Cliques form in citation networks when pairs of papers are commonly cited together.

The meaning of what a community structure represents needs to be determined by the researcher. In social networks, communities may constitute acquaintances or colleagues that have frequent interactions. In a collaboration network, the structure of the co-authorship graph can expose institutional ties or different topics being studied. Communities in biological networks can identify metabolic pathways on the microscopic scale or predator-prey interdependencies on the macroscopic. Due to the efficient nature of being able to traverse across a complex network in relatively few hops, technological networks such as the air transport network or the Internet routing network reflect geographic closeness and population distributions. Trading blocs translate into communities in economic networks. In information networks such as web-pages

on the World Wide Web and citations of academic papers, communities can illuminate which nodes share context and have sections of similar content.

While one of the defining qualities of a complex network is *clustering* and there is a clustering coefficient (given in Equation 3.1), that usage refers more to the small-scale or microscopic behaviour of linking and to avoid confusion in this thesis, clustering will only be used to describe the groupings based on text similarity of documents in the Dark Matter corpus (developed in Chapter 4 and combined with network methods in Chapter 5) and community to describe the mesoscopic groups of nodes on a network.

Data Clustering is a similar concept to community finding and will help to clarify the meaning of community by way of contrast. Clustering is a technique of unsupervised learning that finds groups of objects that share similar properties. The document clustering in Section 4.6 is an example of this. It creates clusters of documents that are textually similar. The clustering process projects objects into an abstract dimensional space representing their properties and groups together objects that are near to each other in that space, given a distance metric. An object's position in that dimensional space is due only to the object's properties in isolation. No interactions between objects are considered.

In a network, community refers only to the nodes it contains as members, but it is the links that make the community. Whether it is to associate with another person in a social network, to activate a metabolic pathway in a biological network or to utilise a fragment from another academic paper in an argument in a citation network, community is defined by choice or activity. Links can represent many relationships. In a social network, they can be links to family or friends, work or social connections. People can belong to more than one community and the strength of their connections is not strictly uniform. In contrast, the references in a citation network are binary and although the purpose of the reference could colour the character of the reference, the choice of making the reference is not arbitrary. There is a good chance that an academic paper and the one that it references share at least one topic, however marginally.

3.3.1 Properties

On the surface, the topic of Dark Matter is made of at least three or four disparate communities: observational astronomy, theoretical astrophysics, high-energy particle physics and the theoreticians studying the alternative model of gravity known as MOND. Prior to collecting the data, it was expected that communities would form around the differing facets of the topic. The act of citation is a flag indicating the author's awareness of other work as well as placing the author's work in the context of current research and acknowledging the relevant literature.

While links outside the community are by definition less common, in social networks at least, these *bridges* between communities are responsible for overall network cohesion. Granovetter's work showed that it is the acquaintances and not the close friends that are the most important sources of information in the job search [Granovetter, 1973]. Equally, bridges provide an efficient transport of information across the whole network [Friedkin, 1982].

Looking at the reverse scenario, Karsai et al. [2014] recently studied a communication network of mobile phone call data using a rumour spreading model. Similar to models of the transmission of infectious diseases, each node can be in one of three states: Ignorant, Spreader or Stifler; with transitions between states being permitted. Their study shows that strong ties actually constrain the dynamic process of information flow across the network. How these ideas of information flow affect knowledge transmission across a citation network remains to be seen.

A side effect of creating communities by assigning each node to a subset of the network is that it may not entirely describe the true community structure of the network. The concept of overlapping communities has been explored by Evans et al. [2011] who re-examined two classic social networks, Zachary's Karate Club and the American College Football network [Zachary, 1977, Girvan and Newman, 2002] by creating a weighted graph of the cliques found in those networks. This allowed the underlying nodes to be assigned a fractional membership. No conclusions were drawn regarding the meaning of the overlapping communities, only that this transformation was an ef-

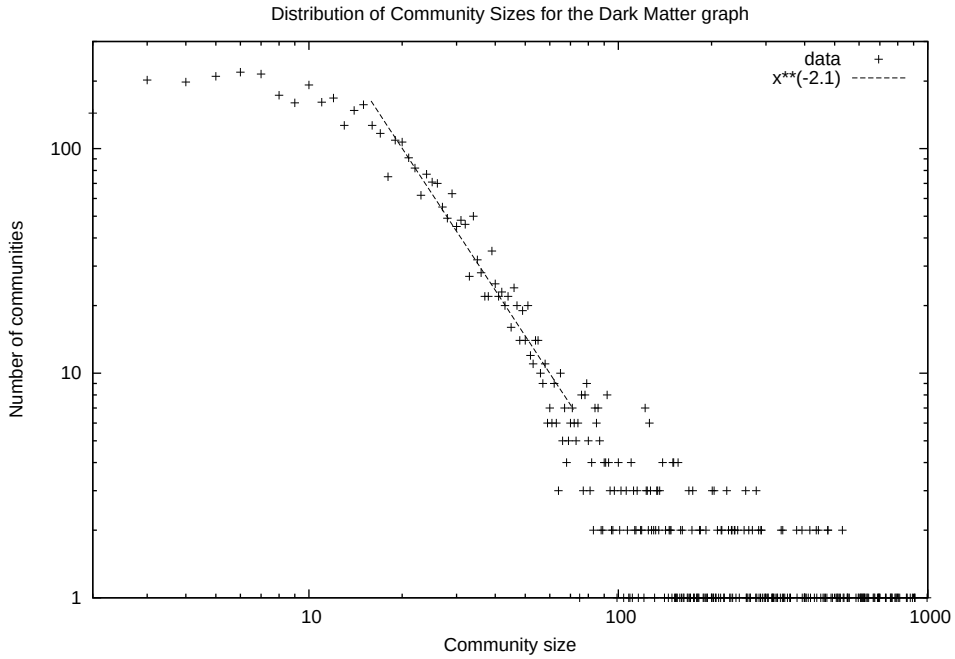

Figure 3.7: Distribution of Community Sizes found by the Louvain algorithm during its first community finding pass. A fit of $x^{-2.1}$ was aligned with the data for comparison.

ficient method of finding them. Identifying community overlap was not attempted for the Dark Matter network, as the preferred choice was to use a crisp set in comparing the similarity of papers.

With respect to the distribution of community sizes, it was found that at the lowest level of the hierarchy, the communities in the Dark Matter citation network follow a power law relation, analogous to the degree distribution. Shown in Figure 3.7, this result has been noted before, but unlike the degree distribution, no clear mechanism for the distribution of community sizes has been devised [Battiston et al., 2007]. Danon et al. [2007] have suggested that it may be caused by the algorithm used to separate the nodes into communities and that it is a feature of the modularity function while Clauset et al. [2004] speculate that it could also be due to the sociology of the network.

3.4 Community Detection Algorithms

Fortunato [2010] published a review on communities in networks that provides an extensive list of algorithms with detailed descriptions as well as commentary on various aspects of communities. It is a comprehensive source and, despite work on algorithms both new and modified being produced steadily [Steen et al., 2011, Schaub et al., 2012, Yueping, 2011], his manuscript will be the definitive reference on the subject for many years to come.

Broken down by approach, it covers:

Traditional methods i.e. the Graph Partitioning Problem, a classic of computer science; and hierarchical clustering, an early technique for social network analysis (SNA) [Scott, 2000].

Divisive methods break the network into smaller equal-sized pieces. The seminal Girvan-Newman algorithm exploits the concept of betweenness centrality [Girvan and Newman, 2002], finding the link involved with the most connections and removing it. When the network finally breaks in two, the process begins again on the two sub-networks. The end result is a binary tree of communities. This has the advantage of being able to inspect the community structure at any scale, but re-computing the betweenness centrality measure after every link removal is computationally very costly.

Spectral algorithms use the algebraic properties of matrices to partition the network. Rather than working with the adjacency matrix whose elements, e_{ij}, represent the number of links between nodes i and j, spectral algorithms construct the Laplacian matrix defined as the element, $e_{ij} = -1$, where there is a link between nodes i and j and the diagonal elements, e_{ii}, are the degree of node i, resulting in a row sum of 0. The Laplacian matrix has the property that a connected network has only one eigenvector with eigenvalue 0. To partition the network, spectral analysis searches for eigenvectors with eigenvalues *close* to 0.

Dynamic methods i.e. Potts or Ising model originally developed for dealing with coupled spin states of atoms [Dorogovtsev and Mendes, 2003] and Random walks as used by Rosvall and Bergstrom [2008] to map out the shape of science.

Within this comprehensive list, it is the topic of *modularity* that he devotes a significant portion of his attention. The modularity function is the sum of weights of links between communities, c, greater than that expected in a randomly-wired network. The modularity, Q, (derived in Newman [2004]) is given by

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \tag{3.2}$$

where A_{ij} represents the weight of the link between nodes i and j, $k_i = \sum_j A_{ij}$, $k_j = \sum_i A_{ij}$, $m = \frac{1}{2} \sum_{ij} A_{ij}$ and δ is a Kronecker delta function selecting only nodes in the same community (1 when $c_i = c_j$, 0 otherwise where node i belongs to community c_i). Given that $k_i k_j / 2m$ is the probability of a link existing between i and j, a non-zero value for Q represents a departure from the random linkages of an equilibrium network. The function is defined in such a way that Q has a maximum value of 1.

The Louvain algorithm developed at the Université Catholique de Louvain (based on the Fast Greedy optimisation of modularity) was chosen for the task of finding communities. It increases the value of Q by combining small communities into larger ones. This makes it a very fast algorithm, and with over a half million nodes, speed is of great concern when analysing the data. The Louvain algorithm iterates over two phases to find the maximum value for Q. It starts with each node having been assigned its own community and then calculates, for each neighbour j of node i, the gain in Q by moving i into a community with j. Blondel et al. [2008] took advantage of the fact the expression of the difference in modularity is quick to calculate, such that very large networks ($> 10^6$ nodes) are only limited by the storage space required by the network, not the computation involved. The difference is given by

$$\Delta Q = \left[\frac{\sum_{in} + k_{i,in}}{2m} - \left(\frac{\sum_{tot} + k_i}{2m} \right)^2 \right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m} \right)^2 - \left(\frac{k_i}{2m} \right)^2 \right] \tag{3.3}$$

where $\sum_{in}$ is the sum of the weights inside the community, $\sum_{tot}$ is the sum of the weights of links to the community, k_i is the sum of the weights of the links to node i and $k_{i,in}$ is the sum of the weights from i to nodes inside the community.

Using the Louvain algorithm to find communities implies that, throughout this work, a community shall be defined not just as a subset of nodes that have a higher density of links to nodes within the subset than to nodes outside the subset, but as a subset of nodes where the density of links within the subset is higher than would be expected from the random linkages (the $k_i k_j / 2m$ term in Equation 3.2) found in a network constructed without community structure.

An additional feature of this method is that it creates partitions of different scales, a hierarchy which has 4 levels in the Dark Matter network. Comparisons with other algorithms have found that only two of the algorithms they tested find better partitions, but are costlier in terms of computation [Blondel et al., 2008]. While Lee and Cunningham [2014] have found that Louvain has issues finding smaller scale communities on large social networks, its performance has seemed adequate for the task. Not insignificant in the decision to run an algorithm is the ease of compiling or installing the package. One researcher selected the Louvain algorithm on the basis that its use was "convenient" [Evans et al., 2011].

Once a partition has been created, communities need to be assigned labels in order for deeper analysis to continue. The act of giving meaning to these groups is not straight forward. Chen and Redner [2010] identified their Physical Review citation network using the Phys. Rev. family of five journals, named Phys. Rev. A through Phys. Rev. E, which splits that once-consolidated publication along broad topics. They have been fortunate that this avenue was available to them. Clauset mentions that using metadata associated with the nodes to compare communities in order to check the veracity of his algorithm [Clauset et al., 2004]. The metadata immediately available originates from the ADS and consists of the keywords (if present in the database), authors, publication date and journal. Radicchi et al. [2004] states that validation is not trivial and that no criterion exists to assess their accuracy. The size of the data set makes this even more uncertain and at the highest level of the hierarchy, the keywords

are unsurprisingly quite disparate given the sizes of the communities.

3.5 Modelling Citation Networks

In Costa et al. [2011], it was shown that many types of networks have models, but they reported that as of 2008 (a decade after Redner's first study), there were none for citation networks in the seven studies on citation networks that they examined.

In 2009, Karrer and Newman [2009] proposed a model based on a directed acyclic graph that reproduced the basic features of citation networks, as the nature of the network evolving through time means obviously that older papers cannot cite newer papers. The KN model, however, assumed an infinitely large graph, so there is a point at which the real world data diverges from the model. They assume that this breakdown point is due to the finite size of the network.

Later that same year, Wu and Holme [2009] extend the KN model to better fit the data using two parameters to control the ageing of papers. Their fit is an improvement on the original KN model, showing a similar shape to their real world example (a set of research papers on high-energy physics from the arXiv.org repository from 1992 to 2003). Unfortunately, their model still under-represents the data, meaning that there are possibly more features to account for. They talk about analysing triangles where paper A cites paper B and paper C, where paper B also cites paper C.

A more theoretical approach is taken by Wu et al. [2014], who have developed a generalised model of Preferential Attachment that focuses on the rate of citation accumulation. While noting a "first-movers" advantage for early papers, the model also provides a "forgetting" mechanism, suggesting that older papers eventually become irrelevant. They validated their model against three domains in the Computer Science literature. Factors in longevity and reasons a paper is remembered or forgotten are discussed in Abt [1998b] as well as the variation between fields in their time-scales. Golosovsky and Solomon [2013] have produced a model for citation networks of non-linear autocatalytic growth which, if true, would replace the scale-free hypothesis for

this type of network. Most papers are shown to have a lifetime of 6–10 years, but the non-linearity of the growth models allows a few papers an almost infinite lifetime. Their model is based on data from the ISI and includes mechanisms at the level of the individual such as citation copying [Golosovsky and Solomon, 2014]. Finally, Jabr-Hamdan et al. [2014] considers the effect of varying the cap on the number of connections allowed per node in the Krapivsky-Redner growth model with what they call "super-joiners".

3.6 Analysis of Dark Matter Citation Network

N	total number of nodes	778 491
L	total number of links	7 606 982
$\bar{k}_{in}$	average in-degree	9.77
$\bar{k}_{out}$	average out-degree	9.77
$\bar{C}$	average clustering coefficient	0.216
γ_i	in degree distribution exponent	2.3
$\langle l \rangle$	mean path length	8.30
δ	diameter (longest shortest path)	11

Table 3.1: Network Statistics

The Dark Matter network from the ADS database consists of 778 491 papers and 7 606 982 citations. This and other statistics of the network are listed in Table 3.1 to allow comparisons with other complex networks such as Redner's citation data from *Physical Review D* and the ISI database [Dorogovtsev and Mendes, 2003, p. 32]. The Dark Matter network is close in size to the ISI database of papers from January 1981 to June 1997. While the average in-degree is similar (9.77 compared with $\bar{k}_i$ = 8.57), the exponent of the power law fitted to the degree distribution, $P(k_i) \propto k_i^{-\gamma_i}$, falls outside the range of values of 2.5 – 3.0 estimated by Redner, Tsallis and Krapivsky [Dorogovtsev and Mendes, 2003, 3 different analyses listed on p. 80]. The distribution from which this exponent is derived was plotted on a double log scale in Figure 3.8 and

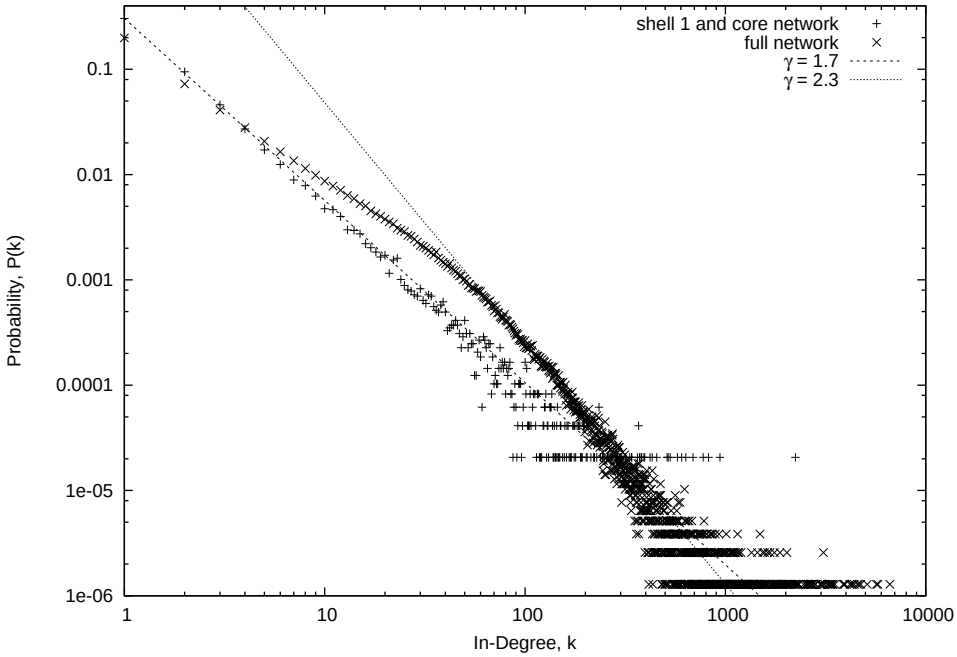

Figure 3.8: In-degree distribution of the Dark Matter citation network, P(k), for the full network and for a smaller network consisting of the core network and shell 1 (described in Section 2.3). The two lines of best fit follow the power law $P(k) \propto k^{-\gamma}$ where the exponent γ is a descriptive statistic for power law relationships.

a fit to the data in the range $k \in [80..300]$ of the form $P(k_i) = Ak_i^{-\gamma_i}$ was produced. For k_i less than 40, the distribution is less steep. Also included in the figure for comparison is the in-degree distribution for the core nodes in the Dark Matter corpus and their immediate neighbours. With only 49 565 nodes in Shell 1, this curve does not have the "knee" of the full citation network and was fitted using a power law with an exponent of 1.7.

In relation to the formal definition for a complex network as given in Section 3.1.3, both distributions are clearly fat-tailed with power law exponents lying within the required range for a complex network. The average clustering coefficient for a random graph with the same number of nodes and links as the Dark Matter network is $C_r = 1.3 \times 10^{-5}$ meaning that there is significant clustering, the ratio of the two co-efficients

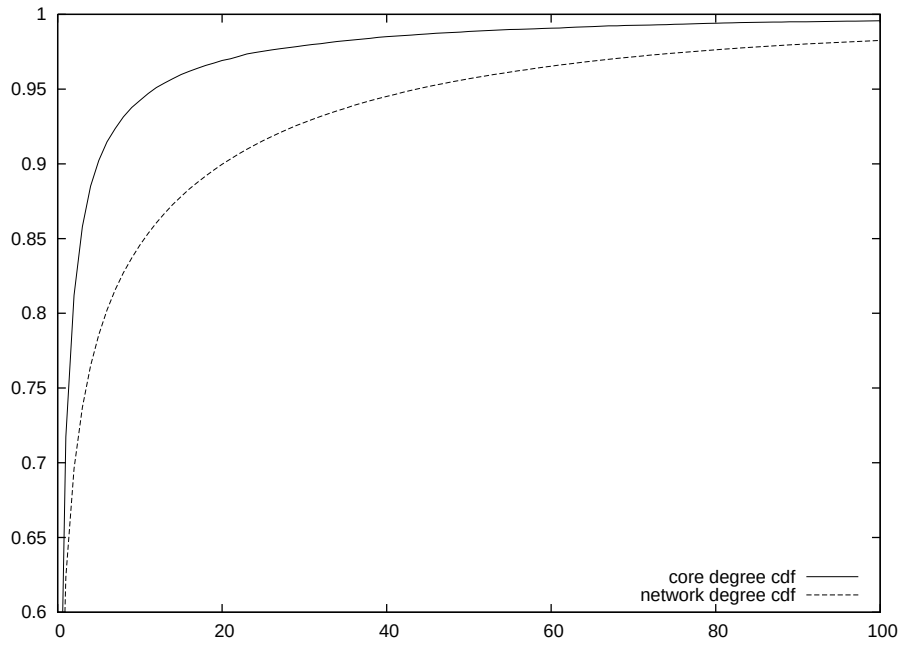

Figure 3.9: CDF of the In-degree distribution—the Dark Matter core network + shell 1 (49 565 nodes) is contrasted with the full network (778 491 nodes)

$\frac{C}{C_r} = 1.7 \times 10^4$, satisfying the clustering requirement for a complex network.

The expected mean path length of a random graph the same size as the Dark Matter network is $\langle l_r \rangle = ln(N)/ln(k) = 5.95$, which is similar in magnitude to the calculated mean path length $\langle l \rangle = 8.30$ for the Dark Matter network. The ratio of the two lengths is $\frac{\langle l \rangle}{\langle l_r \rangle} = 1.39 < 2$, which satisfies the small-world property requirement for the complex network. In addition, the diameter, δ, of the network was found to be 11. This statistic says that the shortest path between any two nodes is at most 11 links, demonstrating the small-world property of complex networks.

To better understand how the degree distribution changes as the network expands, the Cumulative Distribution Function (CDF) of both distributions was plotted in Figure 3.9. Defined as $CDF = \int_0^k P(k)dk$, the CDF shows how the degree distribution shifts in emphasis towards higher degrees with an order of magnitude increase in the number of nodes.

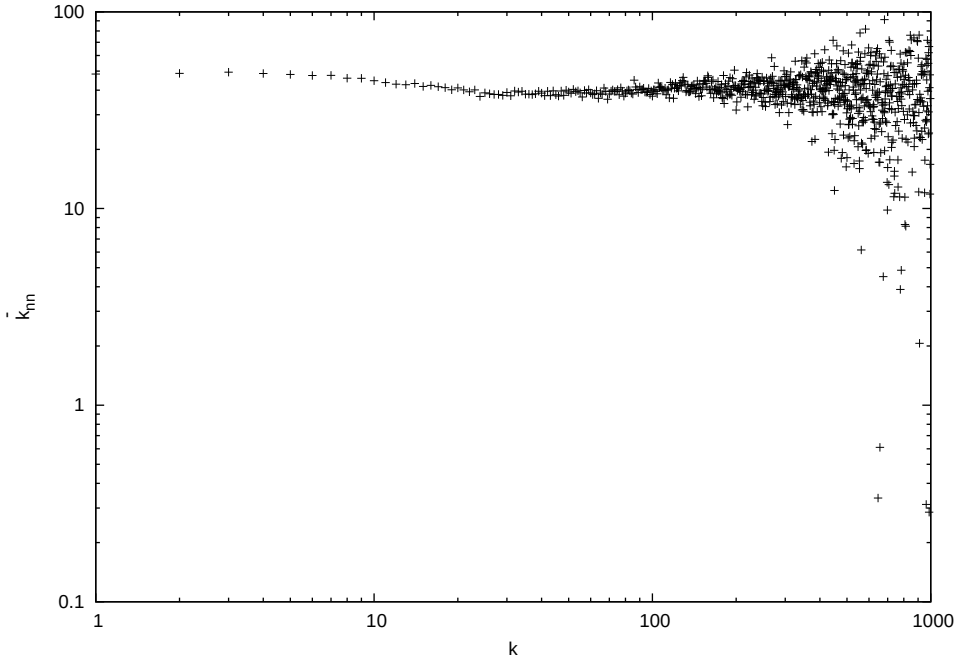

Figure 3.10: Average Pair correlation of citations for Dark Matter citation network, plotting degree, k, against the average degree of its nearest neighbours, $\bar{k}_{nn}$, described in Equation 3.4. An expanded view of the data structure for $k < 100$ is shown in Figure 3.11

To avoid the poor statistics inherent in constructing the joint degree-degree distribution, $P(k, k')$, Pastor-Satorras et al. [2001] used the average degree of the nearest neighbours, $\bar{k}_{nn}$, given by

$$\bar{k}_{nn}(k) = \sum_{k'} k' P(k'|k) \tag{3.4}$$

for each node in an undirected network to study the pair correlations between nodes. Figure 3.10 shows the plot of the average number of citations, $\bar{k}_{i_{nn}}$, possessed by papers of citing papers with k_i citations. While the plot of the whole range (a) shows very little correlation between the two, a close examination of $k < 30$, seen in (b), demonstrates a clear correlation between k_i and $k_{i_{nn}}$ similar to the preferential attachment displayed by the topology of the Internet.

Many different properties exist for complex networks and it is not necessary to

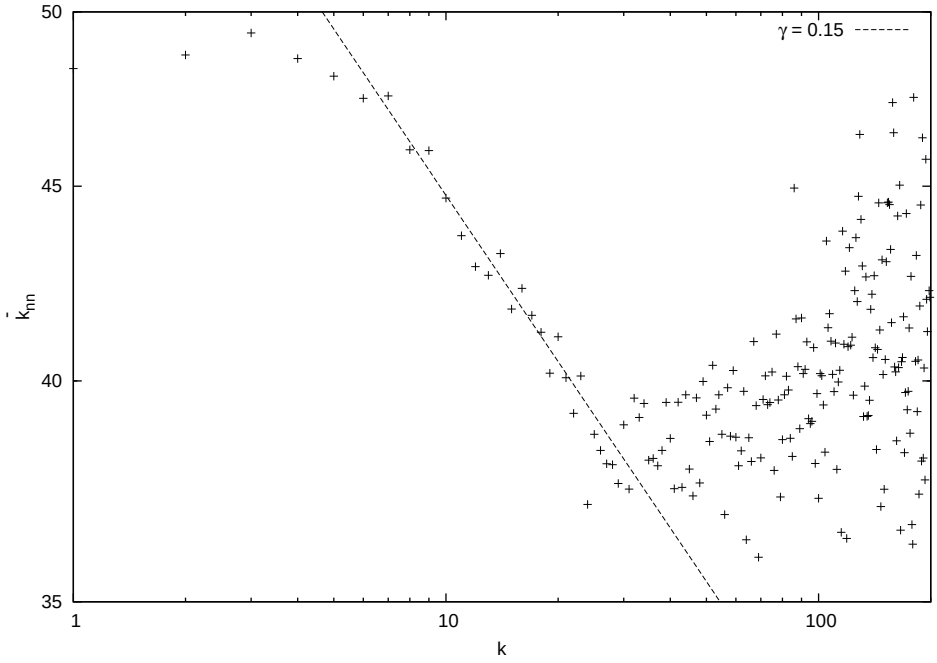

Figure 3.11: Pair correlation at low degree from Figure 3.10 stretched vertically showing linear feature in log-log plot. A fit of $\bar{k}_{nn} \propto k^{-0.15}$ is plotted through the data stretching over the range $k = (7, 30)$.

find all of them in order to compare one network to another. The statistics and distributions presented here should give sufficient basis that the Dark Matter network constructed from the ADS database can be placed in context with other citation networks past and future. Any large discrepancies between them will highlight areas for further investigation.

3.7 Summary

Although the ADS citation database has been used to aid researchers by displaying requested parts of the citation network and clustering keywords for knowledge discovery, its network characteristics have not been analysed. Rivalling Redner's network constructed from the ISI database in size [Dorogovtsev and Mendes, 2003], the degree

distribution was found to be less steep than Redner's. The clustering coefficient was found and its distribution plotted. The distribution of the community sizes found by the Louvain algorithm is similar to other networks, but still lacks an explanation for its behaviour. In the region of low degree, a clear correlation between the degrees of connected nodes was found. This correlation disappeared as the node's degree increased. These findings were compared to previous citation and other complex networks.

4 Text Mining

This chapter introduces Text Mining, a sub-field of Data Mining, and places it in context with several fields that deal with extracting meaning from text. In pursuing their own goals, distinct but related fields such as Information Retrieval, Natural Language Processing and Artificial Intelligence have contributed tools and concepts to Text Mining. Once the output is obtained, it can then be used as input for other tasks. A section on Knowledge Representation is included to explain ontology, a tool used with citations and argumentation, and introduce the concept of representing a document in a form susceptible to manipulation. After outlining the generalised process of Text Mining, the method used here is described and preliminary results reported.

4.1 Introduction

Data mining is the automated or semi-automated process of finding non-trivial and meaningful patterns within a large volume of data. [Fayyad et al., 1996, Witten et al., 2011]. It goes beyond statistical analysis and takes advantage of machine learning techniques to handle large, noisy, messy datasets [Nisbet et al., 2009]. Data mining tasks can consist one or more of the following: exploratory data analysis, descriptive modelling, predictive modelling, pattern discovery or rule finding [Hand et al., 2001].

While Data Mining usually works with structured or numerical data, the related area of Text Mining focuses on the extraction of new information from unstructured textual data [Hearst, 1999, Aggarwal and Zhai, 2012a]. Like Data Mining, it is the automated or semi-automated process of discovering meaningful patterns in text. Text mining has its roots in the study of Information Retrieval (IR) in the 1950s and 1960s [Weiss et al., 2010b, Baeza-Yates and Ribeiro-Neto, 2011], which has the goal of improving access to information. Search engines are a powerful application of IR research, but only present existing documents to the user. Information Retrieval is not traditionally focused on analysing text data for pattern discovery [Jones, 1997, Croft et al., 2010].

Text Mining goes beyond IR in attempting to digest information to support decision making. Information extraction is an important task and is the first step of many applications in Text Mining and other fields [Jiang, 2012]. Early work in the preliminary processing was advanced by the field of Natural Language Processing (NLP), which developed in the 1960s and 1970s. Of interest to both Linguistics and the Artificial Intelligence (AI) community, NLP is concerned with the input and output of natural language, as opposed to formal or artificial languages. Its goal is the understanding of *how* we use language to communicate ideas and is studied on many levels [Jurafsky and Martin, 2008]. Through AI, NLP attempts to make machines understand human communication allowing for improved human-machine interactions, such as providing natural language interfaces to databases [Androutsopoulos et al., 1995]. Natural language has been used in application from classifying documents to filtering spam emails [Segaran, 2007]. With AI researchers interested in NLP, it seems only natural that NLP should borrow machine learning techniques from AI for language modelling [Daelemans and Bosch, 2005], speech recognition, question answering, sentiment analysis and representation of meaning [Abney, 2007]. Such high level understanding tasks proceed through knowledge of linguistics and its structures [Jurafsky and Martin, 2008, Chapter 5]. The hierarchy of linguistic structures from least to most complex would include: phonemes, morphemes, lexemes, syntax, grammar, semantics and pragmatics.

Syntax concerns itself with linguistic expressions only. Semantics is the study of meaning and connects statements to the objects they are expressing. Pragmatics examines the context in which syntax and semantics occur and tries to resolve ambiguities in semantically correct statements by applying knowledge of how the world operates [Montague, 1974]. These linguistic structures are also described by Mary Harris [1985] as she puts into context the developments in linguistics emerging from the work of Franz Boas, Leonard Bloomfield, Zellig Harris, Noam Chomsky, Charlies Filmore, Jerrold Katz and Jerry Fodor. Their research over decades strives for a complete understanding of language, yet we can extract some of the concepts conveyed by simplistic approaches that discard much of the sentence structure. Latent Semantic Indexing (LSI), a high-dimensional linear associative model, derives an implicit rep-

resentation of text semantics from the observed co-occurrence of words [Deerwester et al., 1990, Landauer and Dumais, 1997, Bullinaria and Levy, 2007]. Information Retrieval has made use of the "unigram language model" which discards word order and other contextual information to improve the speed of returning search results [Manning et al., 2008, pp. 221-222]. The Vector Space Model (VSM), commonly used in comparing similarity between documents, accepts the trade-off of less-than-complete meaning in exchange for greatly increased ease of computation [Lee et al., 2005, Mikolov et al., 2013] and is described in Section 4.1.2.

One of the goals of Artificial Intelligence is to create viable machine learning algorithms so that computers can begin to teach themselves. Some applications include assisting with medical diagnoses or image analysis [Ananiadou and McNaught, 2006, Simpson and Demner-Fushman, 2012]. Strategies to achieve machine learning can be divided into supervised and unsupervised algorithms. Supervised algorithms are told what the expected response should be on a training data set while unsupervised algorithms try to discern relevant patterns from the data without external input [Alpaydin, 2004, Nisbet et al., 2009]. It is through these algorithms that the connection between AI and Text Mining is most relevant.

Text Mining tasks include Information Extraction, Text Summarisation as well as supervised learning for Document Classification and unsupervised learning for areas such as Latent Semantic Indexing and Topic Modelling [Aggarwal and Zhai, 2012a]. Classification of documents is usually achieved with a supervised algorithm combined with domain knowledge of the expected topics for such tasks as creating systems that can automatically categorise newly presented documents [Duda et al., 2001, Glover et al., 2002, Aggarwal and Zhai, 2012c]. The clustering of documents is an unsupervised gathering together of similar material into groups and the first step in Topic Modelling [Kaufman and Rousseeuw, 2009, Muresan and Harper, 2004]. Clusters provide the framework from which latent topics are manually identified. The topics serve as reduced dimensions in semantic space. Terms are associated with topics in a probabilistic model of topics in the corpus. Once produced, a topic model can be used to simplify the classification of documents while trying to retain all the significant semantic features.

Topic Modelling's scalability as the corpus expands with new additions, its availability to interpretation by experts and its richer document representation are the advantages that it has over Latent Semantic Indexing. The validity of a topic model needs to be evaluated and is usually undertaken with a specific application under consideration [Wei and Croft, 2006, Wallach et al., 2009]. Some researchers, such as Chang and Blei [2009], have incorporated structural data linking documents into their evaluation on the assumption that topic distributions are similar for linked documents.

Because unsupervised learning can have ambiguous group boundaries, Zhang et al. [2008] and Angelova and Siersdorfer [2006] have both found that links between documents can enhance the quality of the resulting clusters in the face of ambiguity. Zhong [2005] has developed techniques for clustering data streamed along a temporal dimension such as news feeds, and Popescul and Ungar [2000] have investigated methods of automatically labelling clusters. Some researchers combine the two algorithmic approaches in a semi-supervised system which begins with user input and then continues unsupervised [Basu et al., 2004].

One method of providing useful labels for clusters is by identifying the topic or topics that the documents have in common [Stein and Eissen, 2004]. Topic Modelling uses word clusters to reduce the dimensionality of the documents in order to produce document clusters [Hofmann, 1999, McFarland et al., 2013]. This technique has been used on scholarly articles to classify them according to subject [Erosheva et al., 2004] and as the basis for a scholarly recommender system [Wang and Blei, 2011].

The motivation for Text Mining is that with the rapid growth in the number of academic publications calling for attention [Abt, 1998a, Price, 1965], researchers can easily be overwhelmed with information. In the context of social relationships, there is a notional maximum number of relations that the brain can keep track of, called Dunbar's number [Dunbar, 1992]. This maximum extends outside social relationships [Goncalves et al., 2011] and is pertinent to the way researchers approach the literature. There is a need for automated means to process and analyse research literature [Delen and Crossland, 2008]. Text Mining can reduce vast amounts of information to manageable quantities.

4.1.1 Knowledge Representation

Artificial Intelligence is the pursuit of creating machines that can act more like humans and in part through mimicry, seeks to understand aspects of human intelligence. A machine-simulated intelligence requires a representation of the external world. How to represent knowledge is one of AI's major philosophical questions. An ontology is used in AI for knowledge representation as the specification of a conceptualisation. Not to be confused with the same term used in metaphysics to describe categories of being, this usage refers to an agreed vocabulary enabling practitioners to describe ideas with precision and to encode them in machine learning algorithms. Commonly these take the form of hierarchies of classes of the concepts, showing the relationship between instances and are used in rule-based expert systems and agent-based systems to structure the domain knowledge [Callan, 2003]. Ontologies have been used to integrate background knowledge into clustering tasks [Hotho et al., 2003].

One relevant ontology is CiTO, the Citation Typing Ontology, created to describe the actions and properties of academic citations.

> Citations are described in terms of the factual and rhetorical relationships between citing publication and cited publication, the in-text and global citation frequencies of each cited work, and the nature of the cited work itself, including its publication and peer review status. [Shotton, 2010]

The standardisation of citation metadata was designed to bring the commentary method of classical and biblical scholarship into biomedical research practice. By representing knowledge in a machine-readable form, it becomes possible for software to utilise the encoded information in applications from tools to aid researchers assess the strength and nature of scholarly claims to taking its place in the "Semantic Web" [Ceravolo et al., 2006].

The Virtual Observatory project being developed by the Harvard-Smithsonian Institute for Astrophysics is using CiTO in conjunction with the "Provisioning, Authoring and Versioning" ontology and others to describe the relationships between Astronomical Objects, Observatories and Publications as it relates to the Research Lifecycle in Astronomy [Accomazzi and Dave, 2011]. Both of these examples are writ-

ten in OWL, the Ontology Web Language, which is defined in a formal language called RDF (Resource Description Framework) concepts such as evidence, authority, credits and critiques [McGuinness, 2004, Antoniou and Harmelen, 2009]. Tackling one facet of Semantic Publishing, it has been integrated into the SWAN Discourse Relationships Ontology and has been used to create a Supporting Claims tool-tip (a link that displays its citation relationship when the mouse hovers over the Claim).

In contrast to textual descriptions, Skupin has visualised the relationships within a set of knowledge domains by mapping them on to a 2-dimensional space [Skupin, 2009]. Using a Self-Organising Map (or SOM) and the abstracts from the Annual Meeting of the Association of American Geographers, he created a landscape within which authors could be located, demonstrating an inter-disciplinary tendency of some who publish across domain boundaries.

4.1.2 The Vector Space Model for Documents

A representation is needed for documents in order to efficiently manipulate items in the corpus. Any representation used in quantifying documents must have a method for calculating a value for the similarity between two documents [Salton et al., 1975]. The calculation should be quick and single-valued and the storage required should be small. Given a corpus of N documents, there will be $N(N-1)$ calculations to determine the distribution of the similarities and for unsupervised learning of the documents into k clusters, the k-means algorithm, described in Section 4.6 and Appendix C, calculates $k \times N$ similarities for *every* iteration until it converges or stops. The use of vectors achieves all of the above through a transformation from textual to numerical representation.

In the most general sense, a vector is a mathematical object that can be added and can be multiplied by a scalar (a number). A vector in this sense is an element of a vector space, in which two vectors added together produce another vector in the set and multiplication of a vector by a scalar also produces another vector in the set [Daintith and Nelson, 1989]. By considering each term in a document as an individual dimension

of a vector space, a document can be represented as a vector whose dimensions are measured by the *term frequency*, which is the number of occurrences of each term found in that document. The term *bag-of-words* originates from remarks by Zellig Harris [1954] on how language consists of more than just unordered words. VSM is sometimes referred to as a bag-of-words representation because the information coded in the order of the words is lost in the process of calculating the similarity [Manning et al., 2008, Ch. 6].

The Vector Space Model (VSM) cuts the Gordian knot of NLP, turning disadvantage to advantage. It side-steps the understanding of syntax and grammar and with it, the associated difficulty in constructing such a system. It certainly loses the relationships between actors, events and locations, but it is not troubled by ambiguity in meaning. In fact, it treats lexical ambiguity as a composite meaning which has the advantage of being consistent. As both classification and unsupervised learning are concerned with grouping objects that are alike, precision is not essential because our interest is in the group as a whole, not the detail of the individual. The model is simple, easy to implement, quick to compute and compact to store. The return on the effort to extract meaning computationally is similar to the Pareto principle in that VSM retains a large portion of the meaning of a text with only a fraction of the effort required to fully parse the text. A study shows that 80% of information comes from word choice while 20% comes from word order [Landauer, 2002]. One inherent disadvantage that VSM shares with NLP is that it can only account for what is explicitly written. Assumed knowledge is not represented.

4.2 Document Selection

Document Selection is the first step in a text mining process. In this application, it reduces to the simple choice of an individual paper being the document unit (perhaps with more than one source file which could be an issue for some papers). Because the documents are transformed to XML, accented character sequences are converted

to Unicode, providing uniform treatment across the collection. The other advantage of using a parser that produces XML is the number of parsing options available. For tokenising (see Section 4.3.1), a stream-based parser is faster, uses less memory and is perfectly suited for identifying "chunks" of data. In attempting to identify sections of text according to their nearby citations, it would be preferable to use a tree-based XML parser to perform the backwards searching required as long as the entire document representation could be held in memory.

4.3 Parsing

4.3.1 Tokenising

The first step in extracting text from a document is to break the text up into pieces called *tokens*, which are character sequences passed along to the next stage in processing. At this stage, it is convenient to remove unwanted characters, such as punctuation marks, to ease the following processing steps. Initial options for converting LaTeX to plain text were `dvi2tty`, `crudetype`, `catdvi` and `latex2text`. A common problem amongst them is the hyphenation that TeX inserts when typesetting at the line breaks. These extra hyphens would have to be removed in order to recover the original words.

The program used is LaTeXML which produces an XML representation of the latex document which can have an *abstract* tag and a *keywords* tag if they have been included in the original document. The abstract and keywords are subsets of the vocabulary model. A related program, `latexmlpost` (see B.1), will produce MathML for the equations. Post-processing also has the option to match the citations to their position in the text, allowing for the possibility of testing claims against the references provided, discussed briefly in Section 7.1.2.

4.3.2 Stop Words

Another technique for improving the semantic impact of the document is by removing words common to all documents [Lee et al., 2005]. These are called stop words and they are the everyday words that glue parts of speech together, but are too general to identify the subject of a document. Witten et al. [2011] calls this a feature selection problem that can affect the calculation by including semantically non-selective terms. The general strategy in Manning et al. [2008] is to sort all the terms by collection frequency and make a judgement as to how many of the top terms to "stop". Many stop word lists exist, usually to speed up search engine queries, although in some applications they have had a detrimental effect on performance [Bullinaria and Levy, 2007].

A deep inspection of stop words should consider capitalisation. When some words are capitalised mid-sentence, their significance changes and so too does the decision regarding their inclusion in the stop word list. The ADS has 4 stop lists; two of them case-sensitive. The treatment in these two cases is that the words to be "stopped" are removed before the text is converted to lowercase. An improved treatment would identify capitalised instances that begin sentences and treat them as lowercase forms.

While care in refining the document representations should improve the similarity calculations as well as reducing the number of dimensions on which to compute document lengths and dot products, which according to Manning is not crucial to success.

One final consideration is to never remove a keyword (an index term that has significant meaning to the topic). These have been chosen as the words that best speak to the meaning and intent of the document. As such, they are highly significant and *must* be included in the analysis.

4.3.3 Normalisation

Normalisation is the process of equating superficially different word forms of the same term, such as anti-hallucinatory and antihallucinatory. This process can be manually extended to synonyms such as car and automobile [Pennell and Liu, 2014]. Initially sets of rules for replacements, some work has gone into normalising non-standard words, such as abbreviations, in an unsupervised manner [Sproat et al., 2001].

The approach used for finding similar word forms was to sort the tokens alphabetically and visually identify the word equivalences. To do this algorithmically, using the edit distance between words to create a shortened list of candidates for verification as equivalents. To find synonyms, one strategy would be to start with the most semantically dense texts, the keywords, then the abstracts, looking for functional equivalents that do not have similar forms.

To better understand the effect of an incomplete normalisation, I calculated the error induced by one word having two forms, finding the distance between two hypothetical document vectors which are identical except for one term. Consider the case of a generalised vector representation of a document where one term has two semantically equivalent forms. Let the vector of the normalised document have n terms with term-frequencies f_i and take the un-normalised representation to be the nth term split between two forms with frequencies, $f_n - g$ and g.

It is shown in Appendix D that the error is less than

$$\frac{g^2 \sum_{i=1}^{n} f_i^2}{(\sum_{i=1}^{n} f_i^2 - f_n g)^2}$$

For any sufficiently large document, the error is less than one percent.

4.3.4 Stemming

Similar to normalisation in Section 4.3.3, stemming addresses the near-equivalence of words that differ by the ending of the word forms, such as the tense of the verb which could end in -ing, -ed or -s or the number of a noun which can be singular or plural.

> The goal of both stemming and (its related process), lemmatisation, is to reduce inflectional forms and sometimes derivationally related forms of a word to a common form. [Manning et al., 2008, p. 30]

The meaning is retained even when small differences, which are significant only in machine representation, are removed. Porter's stemming algorithm is a common method used to remove the suffixes, leaving only the word stems [Porter, 1980].

Porter's algorithm is very important in a postings or index based Information Retrieval system, but is less critical in creating a VSM representation by the same reasoning as with normalisation. Stemming improves the results somewhat by reducing the noise compared to un-processed data and is sufficient for comparing documents.

The Perl module Lingua::Stem::En was used to perform the stemming [Richardson, 1999]. It implements Porter's stemming algorithm and has been acknowledged by Martin Porter on his web page.

Lemmatisation could be considered the rigorous implementation of stemming as it identifies the morphological root of the word. Manning does discuss the limitations of lemmatisation with respect to search, but as only the most common words in English have convoluted expressions, they are the most likely to be removed with the stop words and are unlikely to have a significant impact [Manning et al., 2008]. For this reason, lemmatisation was not used.

4.3.5 Discussion on Parsing

It is possible to produce a VSM representation of a document without removing stop words, stemming or normalising the tokens and compute similarities between documents. Failure to do so adds noise to the metric and makes some documents more dissimilar than others. It can mislead the unsupervised learning algorithm by drawing attention away from significant features.

4.4 Computing Similarity using the Vector Space Model

4.4.1 Using Cosine as a Measure of Similarity

A good distance function or metric can improve the performance of machine learning tasks such as unsupervised learning [Chang et al., 2014]. In linear algebra, the *dot product* of two vectors is equal to the product of their magnitudes (length) and the cosine of the angle between the vectors, as

$$\vec{a} \cdot \vec{b} = \mid \vec{a} \mid\mid \vec{b} \mid \cos\theta \tag{4.1}$$

In this manner, the *cosine similarity* of the vector representations $\vec{V}(d_1)$ and $\vec{V}(d_2)$ is the dot product divided by the magnitudes of the vectors.

$$sim(d_1, d_2) = \frac{\vec{V}(d_1) \cdot \vec{V}(d_2)}{\mid \vec{V}(d_1) \mid\mid \vec{V}(d_2) \mid} \tag{4.2}$$

As each term in the VSM is considered a dimension, the method to calculate the dot product of two document vectors is to sum the product of the number of occurrences in each document for every term, t, given as $\vec{V_1} \cdot \vec{V_2} = \sum_t tf_1 * tf_2$. This function serves as a metric. It is the foundation of the metric space in which all the similarities have a well-defined distance between all other similarities. As all components of the document vectors take non-negative values in the VSM representation and the value of the cosine has a maximum value of 1, all similarities calculated in this fashion must lie in the range $[0, 1]$.

4.4.2 Comparison of Vector Space Model variants

There are a number of weighting schemes to compensate for anomalies in how the VSM is calculated. Manning gives a description of the SMART notation for recognised variants [Manning et al., 2008, Ch. 6.4]. The division by the magnitude of each document

has the advantage of normalising documents to remove the effect of document length from the similarity calculation [Aggarwal and Zhai, 2012b] (not to be confused with Normalising the text, as in Section 4.3.3). The use of the Euclidean length (referred to as cosine normalisation by Manning) to compare documents is sufficient in our case as all documents are expected to be of the same order of magnitude, with the exception of review articles which are few and rarely submitted to arXiv. In the current publishing climate in science, it would be expected that articles that are too long will be broken up by the author into two papers and those that are too short will not be accepted for publication. Using this reasoning, few papers would be more than twice as long as any other, no content is repeated and is devoted to single topics. All documents are equally relevant and of similar content. For these reasons, a *pivoted document length normalisation* is unlikely to be needed. Manning describes a number of weighting schemes that try to discern significant differences between documents, encapsulated by the SMART notation. Based on the inverse document frequency (*idf*), the weights are a variant on dividing the term frequency (*tf*) by the logarithm of the number of documents that a term is found in. This weighting favours terms that are not common across the document collection in a similar manner to the removal of stop words (Section 4.3.2). This is a linear transformation of the vector space, stretching and squeezing the vectors, undergoing change but still consistently measuring similarity. The *tf-idf* method requires all documents to be processed before the similarity calculation can begin. Typically defined as

$$tf\text{-}idf = f_{ij} \; log\left(\frac{N}{df_i}\right) \tag{4.3}$$

where N is the number of documents and df is the number of documents that include word i [Manning et al., 2008]. The CLAIR library (described in Appendix C) uses *tf-idf* when it finds cosine similarities.

It is reasonable to dismiss any concerns over hapax legomena because the only effect a singular occurrence of a word in the corpus has is to increase the length of the document by 1, which decreases the similarity by a factor less than the inverse square of the document length, which is a very small amount for these documents.

The deficits of using VSM are covered by Oliva et al. [2011] in their description of SyMSS, a system measuring semantic similarity. They are correct in saying that VSM loses semantic information, such as treating positive and negative statements as the same. This conflation is not a pitfall for unsupervised learning where groups speaking about the same subject are desired, whether or not they agree on the matter. More intricate investigations exploring topic models have used sophisticated techniques such as Latent Dirichlet Allocation (LDA) and its derivatives [Ramage et al., 2011, Jardine and Teufel, 2014]. Their power comes at the cost of computation and scalability can be an issue. In contrast, VSM is fast and easy to compute.

Similar to the cosine similarity are the Jaccard similarity and the related Sørensen similarity. These measures are binary in nature and only examine the existence of a word in document, not its frequency.

Definition for the Jaccard similarity

$$J(X,Y) = \frac{\mid X \cap Y \mid}{\mid X \cup Y \mid} \tag{4.4}$$

where the Jaccard distance is

$$1 - J(X,Y) \tag{4.5}$$

In contrast to the cosine similarity, the Jaccard similarity does not distinguish between a document that focus heavily on a topic and one that only mentions it in passing. Lee et al. [2005] shows that the cosine and Jaccard similarities are nearly equivalent with the cosine similarity being marginally closer to human judgements of similarity.

4.5 Method

The Dark Matter corpus was downloaded from arXiv as described in Section 3.2.1 and unpacked. Most files were LaTeX source files, but there was a mixture of PDF, PostScript, HTML, and MicroSoft Word files. A few needed manual intervention before parsing could begin. Non-LaTeX files were converted to text using the Linux utilities

`pdftotext`, `ps2ascii` and `lynx`. There were only 2 Word files (less than 0.1% of the corpus), so they were excluded from the processing pipeline.

LaTeX source files were converted using code from the LaTeXML project, described in Appendix B. The command `latexml --nocomments` was used to process the LaTeX files into XML. The **abstract, text** and **p** elements of the resulting XML file were extracted using `XML::Parser`. Text for each was split at the boundaries between alphanumeric and non-alphanumeric characters and changed the resulting tokens to all-lowercase. Each set of tokens from the abstract, keywords and text body was stored in a Perl object created to contain the list of words and the properties to be associated with each vector.

Stop words were removed from the word list using a list of 213 words provided by the Perl module `Lingua::EN::StopWords` combined with 41 words selected from the top 200 most common words found from the corpus. The stop lists as well as the exceptions can be found in Appendix F. The words were then stemmed with the Perl implementation of Porter's algorithm, `Lingua::Stem::En`. The list of words excepted from the stemmer was determined by inspection of the top 200 terms from the corpus. The number of occurrences of each remaining word in the document was counted and written to a word frequency file.

After each document was processed in this manner, each word frequency file was read and the similarity between each document was calculated using the cosine similarity (section 4.4.1) and recorded in a database. There were 2659 papers processed (see Table 2.1). The resulting VSM files ranged in size from 12 to 2230 terms. The mean number of terms was 511 with a standard deviation of 223. The percentiles of the number of terms for files in the Dark Matter corpus are tabulated in Table 4.1 and show a distribution skewed towards higher values with the largest vector being twice as large as the 98th percentile.

percentile	*number of terms*	
2	160	
10	269	
25	355	first quartile
50	475	median
75	643	third quartile
90	791	
98	1071	

Table 4.1: Percentile values of the number of unique terms in each document for the Dark Matter corpus after processing into VSM files

4.6 Clustering

The goal of Text Mining is finding structure in unstructured data [Weiss et al., 2010a]. The purpose of unsupervised learning is to group together objects that share features. This can be the sole reason for the activity or the first step in the Knowledge Discovery procedure [Hashimi et al., 2015]. A good set of clusters are ones in which objects within the same group are similar while objects in different groups are different [Jain and Dubes, 1988, Steinbach et al., 2004]. In Text Mining, clustering is used for such tasks as recommendation systems [Han and Kim, 2016, Kim and Chen, 2015], Information Retrieval [Mishra et al., 2015], Topic Modelling [McFarland et al., 2013], Text Summarisation [Choudhary et al., 2009, Nenkova and McKeown, 2012], forensic analysis of text [Nassif and Hruschka, 2013] and finding gaps and priorities in scientific research [Rabiei et al., 2016]. The power of clustering is in the revelation of hidden and interesting features within the dataset.

The k-means algorithm was conceived in 1955 and has become one of the standard general-purpose clustering tools available to data analysts [Jain, 2010]. k-means has become the baseline algorithm by which many researchers compare new algorithms they have developed [Steinbach et al., 2000, Forsati et al., 2013, Jain and Grewal, 2016] and

a starting point from which to add refinements [Nalawade et al., 2016, Gupta and Srivastava, 2014, Wu et al., 2015, Kaur and Rashid, 2016, Curtin, 2016]. Measures for validating clusters using k-means have been studied in depth [Wu et al., 2009]. Jain and Grewal [2016] found that in comparisons based on outliers, k-means outperformed the density-based approach of DBSCAN [Rehman et al., 2014] in terms of precision, recall, F-measure and efficiency while being slightly poorer in terms of accuracy. In Text Mining, k-means is usually applied to document collections using similarity metrics. Both k-means and the related k-medoids are two of the most widely used distance-based partitioning algorithms [Kaufman and Rousseeuw, 2009, Aggarwal and Zhai, 2012b]. k-means has enjoyed such longevity in part because it is a more efficient means of clustering large document collections [Steinbach et al., 2000, Dhillon and Modha, 2001, Wu et al., 2007].

A drawback of using k-means with the Vector Space Model is the "curse of dimensionality", in which algorithms that work well for low-dimensional objects perform poorly on high-dimensional objects [Bellman, 1961]. The Scatter/Gather algorithm [Cutting et al., 1992] uses a computationally-expensive hierarchical algorithm to find good seed objects to feed into k-means. In a comparison of six algorithms, Nassif and Hruschka [2013] found that k-means and k-medoids can find very good quality clusters if the algorithms are well initialised. A more elegant alternative is to simply truncate the cluster centroids to reduce the dimensionality of the similarity calculations. Schütze and Silverstein [1997] find that, for a modest truncation, this method retains the efficiency of k-means without sacrificing cluster quality. k-means and related methods of grouping individuals into clusters of similar characteristics are examples of hard-clustering algorithms. This type of clustering allocates each member to only one cluster.

tf-idf is a method of improving sensitivity to important features in a bag-of-words model, such as VSM. By weighting the term frequency by the rarity of the term in the corpus, as given in Equation 4.3, it suppresses more common, but less relevant or significant terms. Removing stopwords immediately reduces the dimensional space by eliminating terms common—and therefore less discriminatory—to many of the

documents in the corpus. Both techniques increase the impact of the remaining data and make clusters more distinct.

Inflexible groupings can be awkward in boundary cases where a member could rightly be considered a member of two groups that it sits between. Soft-clustering algorithms such as Expectation-Maximisation (EM) assign to each member probabilities of belonging to a cluster and can handle multiple group memberships. It does this by assuming that each cluster has a probability distribution (usually Gaussian), calculating the probability of each member belonging to those clusters and assigning members to those clusters. It then re-calculates the parameters it has in order to maximise the probabilities that these members belong to it. It iterates through the re-assignment of members to clusters and the re-calculation of cluster parameters until cluster memberships do not change. Like k-means, it is not guaranteed to converge to the global maximum and should be repeated a number of times with different starting values for the parameters. A survey of clustering methods can be found in Berkhin et al. [2006] while Duda et al. [2001] details k-means and the EM algorithms with regards to clustering in pattern recognition.

4.7 Results

There are a number of possibilities for measuring the quality of the clusters. F-measure is a binary classification appropriate for assessing test outcomes in terms of true and false positives and true and false negatives [Rijsbergen, 1979]. The Davies-Bouldin Index measures the separation between clusters in terms of their centroids [Davies and Bouldin, 1979]. The Dunn Index is a method for defining the size of the clusters [Dunn, 1974]. The two most appropriate methods are the Rand Index [Hubert and Arabie, 1985] and Mutual Information [Vinh et al., 2010] because they assess the quality of cluster membership. The Rand Index calculates the ratio of the number of cluster membership agreements to the total number of pairs of elements. Given two partitions of a set S with s elements, $X = \{X_1, \ldots, X_{|X|}\}$ and $Y = \{Y_1, \ldots, Y_{|Y|}\}$, a pair of

elements (e_i, e_j) are in agreement if

$$(e_i, e_j) \in S \mid (e_i, e_j) \in X_k, (e_i, e_j) \in Y_l \tag{4.6}$$

or

$$(e_i, e_j) \in S \mid e_i \in X_k, e_j \in X_m, e_i \in Y_l, e_j \in Y_n \tag{4.7}$$

where $k \neq m$ and $l \neq n$ and $1 \leq k, l, m, n \leq s$. The total number of pairs is given by the binomial coefficient,

$$\begin{pmatrix} s \\ 2 \end{pmatrix} = \frac{s!}{2!(s-2)!}$$

The Rand Index, R, is given by

$$R = \frac{a+b}{a+b+c+d} \tag{4.8}$$

where a is the number of elements in Equation 4.6, b is the number of elements in Equation 4.7 and the denominator $a+b+c+d$ is the binomial coefficient. Note that this statistic is only appropriate for evaluating hard unsupervised learning algorithms such as k-means. A soft clustering algorithm, such as Expectation-Maximisation, produces partial memberships and would require a technique such as the fuzzy extension to the Rand Index developed in Campello [2007].

4.7.1 Distributions

The distribution of cosine similarities between different text structures of the documents, the paper abstract, the keywords and the main text body were plotted together in Figure 4.1 for comparison. These represent over 2 million measurements (157 000 for keywords) which have been placed in 100 bins, counted and expressed as a proportion of the total number of measurements. The textbody and abstract similarities are smooth curves skewed to the smaller values with peaks at 0.22 and 0.08. While the average similarity is low there are a number of documents that share significant amounts of vocabulary. The smoothness could be attributed to the wide spectrum of terms used to express the ideas smearing out the distribution. The average similarity

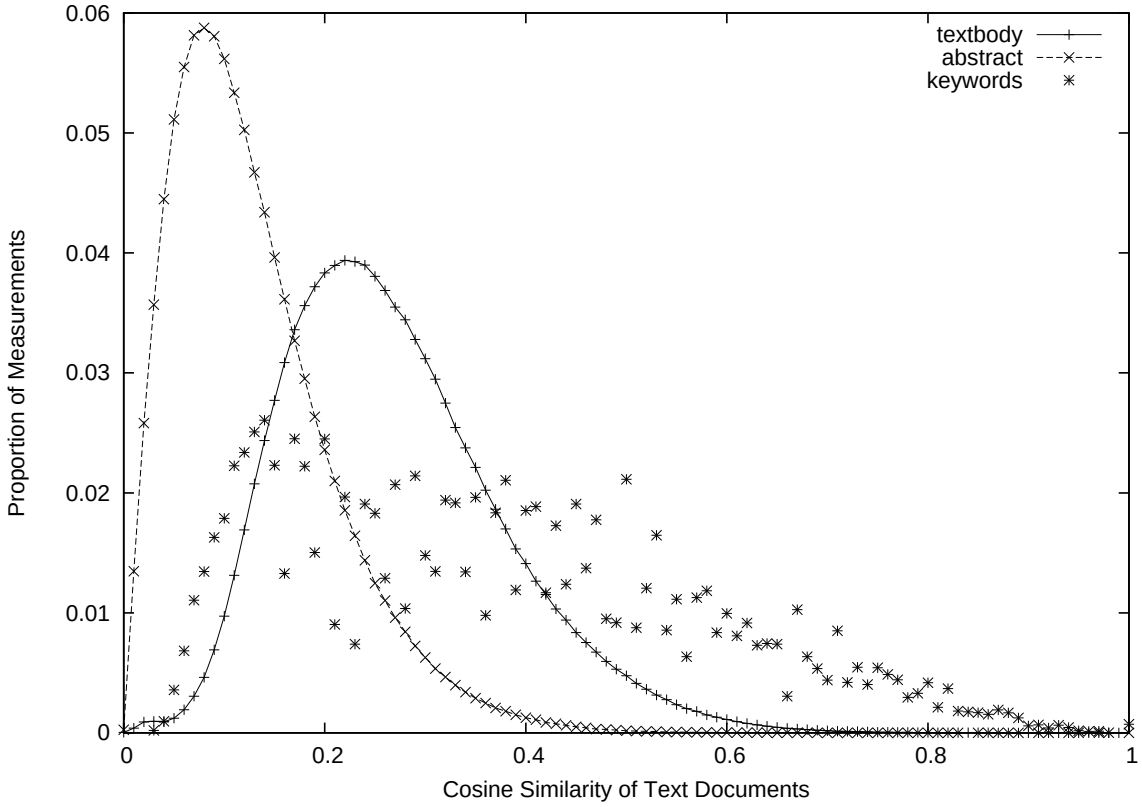

Figure 4.1: Cosine similarity distribution between all text documents in the Dark Matter corpus by document internal structure (abstract, keywords and main textbody).

of the abstracts being roughly half of that for the full text of the paper is most likely due to fewer opportunities in which to find similarities. The platykurtic nature of the keyword distribution suggests the appropriate use of multiple keywords to describe a paper's content, although with only 5% of the measurements of the other document features, the scattered distribution is difficult to interpret.

4.7.2 Similarity over Time

To understand how the passage of time affects document similarity, the similarity was plotted against the difference in time between the published dates of the two documents

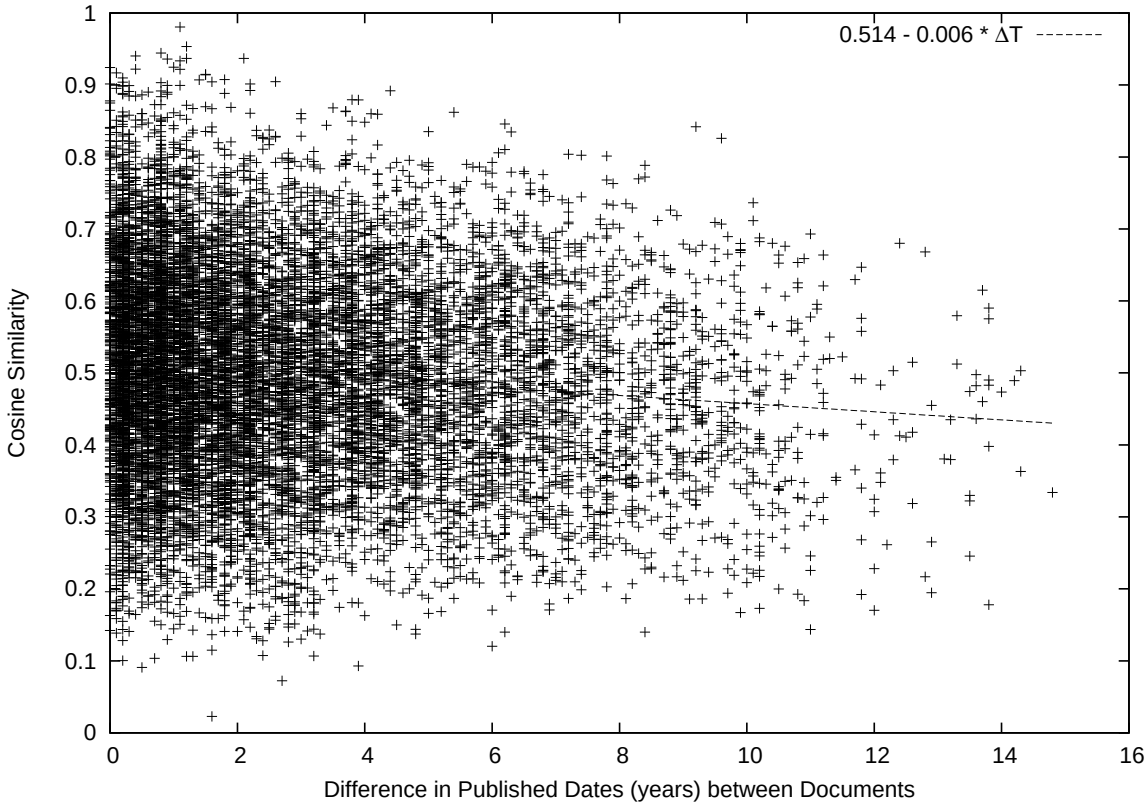

Figure 4.2: Text Similarity over Time. Cosine similarity value of documents linked by citation are plotted against the difference in years between the publication of the two documents. A linear fit to the data was made to show the slow decay in text similarity over time.

and a straight line was fitted to the over 2.6 million data points. A straight line fitted to the data showed that, over a 10-year period, the average document similarity changed by about 1%. The plot was further limited to only those pairs of documents which are connected in the citation network. Here too is very little effect, as seen in Figure 4.2 in which there is only a slight decrease in similarity over time, at the level of 0.06 per decade. A clearer representation of the data is presented as a boxplot in Figure 4.3, in which the line of best fit shows that the median similarity for each year decreases linearly while the statistics are good. To put these values into context, the formula for incomplete normalisation was applied to five small files at the 10^{th} percentile for size

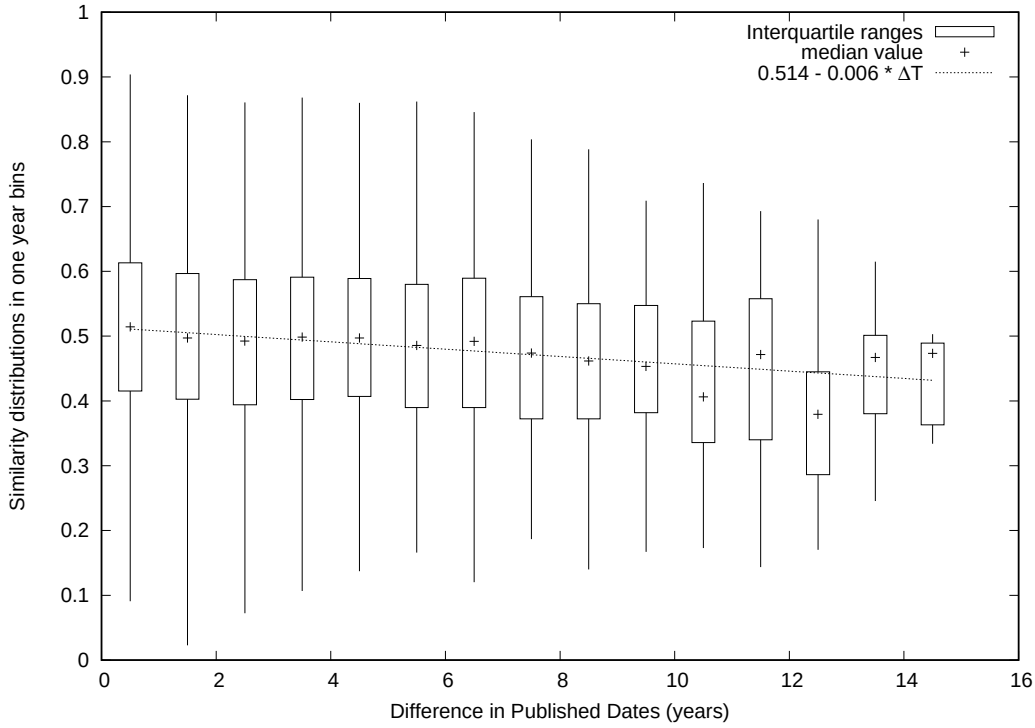

Figure 4.3: Text Similarity distributions over Time. Collected into one year bins and the median values marked with a +, the boxes extend from the first quartile to the third quartile and the lines mark the full range from lowest to highest values. The line of best fit to the data is plotted showing a slope of decreasing similarity value of -0.006/year.

of the Dark Matter corpus (269 unique terms each). Assuming an equal dissimilarity in a term of frequency 4, (putting $f_n = 4, g = 2$ into Equation 4.9), the average upper bound on the error in the similarity value is 0.001, 8.5% of the annual change for the average similarity in Figure 4.2. It would be tenuous to assert that this decrease is due to the shift in research interest over time and could easily be the normal variation in language. Further investigation with other collections of documents, specifically a comparison with a corpus whose topic does not evolve, would be needed to support or reject this claim.

4.7.3 Residual Sum of Squares

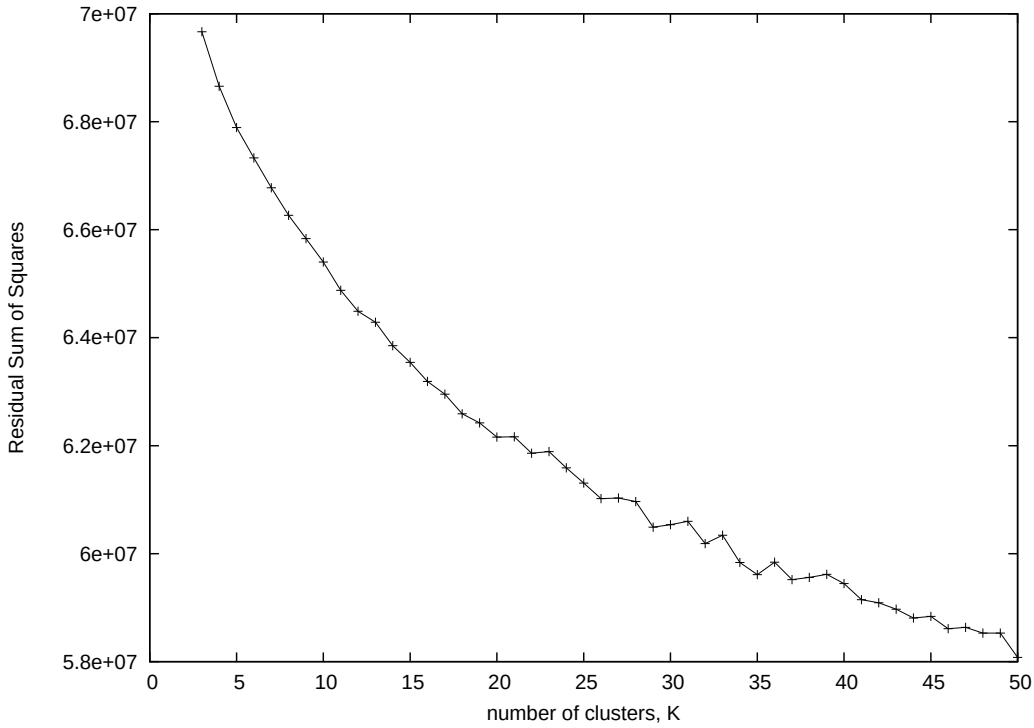

Figure 4.4: Residual Sum of Squares $\widehat{RSS}_{min}$ for k-means clustering of the textbody section of documents in the Dark Matter corpus as a measurement of closeness of cluster members to their k cluster centroids.

The k-means unsupervised learning algorithm was run 10 times for each cluster size from 3 to 50 clusters and the smallest of the total Residual Sum of Squares (the sum of the errors squared between the k cluster centroids and their cluster members, denoted by $\widehat{RSS}_{min}$) was plotted versus the cluster number in Figure 4.4. In Manning et al. [2008, 16.4.1], a "knee" in the line (a flattening of the curve) indicates a natural partition of papers in vocabulary. Unfortunately, there is no clear cluster size indicated. The first flattening of the curve near k = 20 in Figure 4.4 could be seen as just the first signs of noise in an otherwise smooth curve. Zooming in at small k (number of text clusters) in Figure 4.5 and plotting a straight line through two of the points suggest

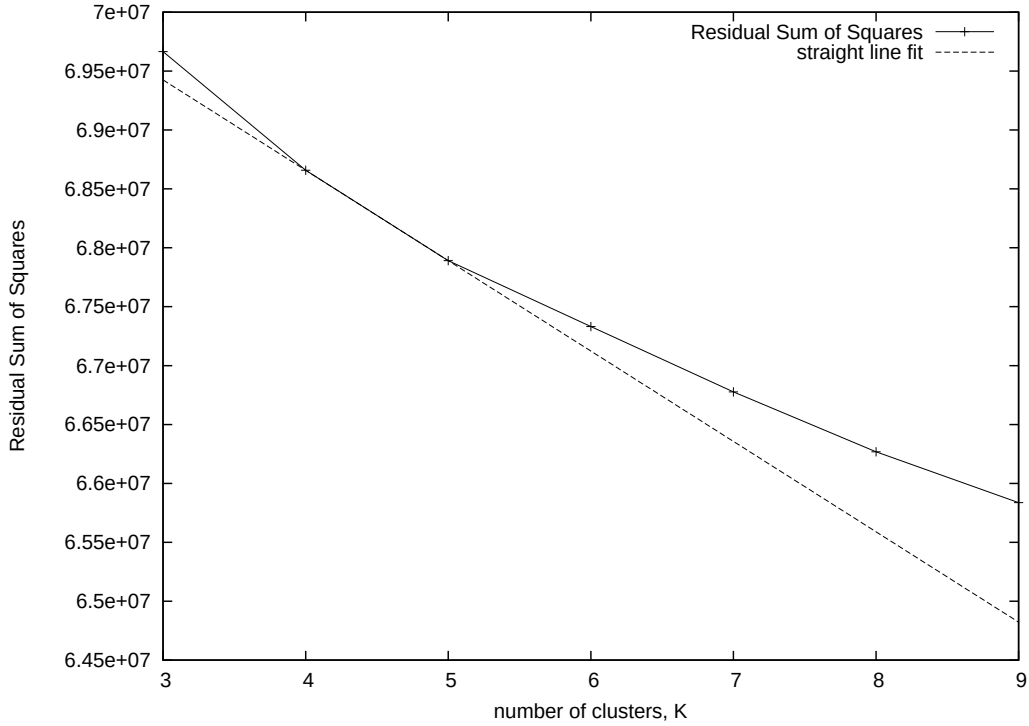

Figure 4.5: Curve of $\widehat{RSS}_{min}$ at small k compared with a straight line, highlighting the change in $\widehat{RSS}_{min}$ at $k = 4, 5$ and 8 in Figure 4.4. Large deviation from a straight line signals a value for k which produces good clusters.

that arguments could be made for cardinality of clusters at any of $k = 4, 5, 8$. Possible reasons for this indiscernibility is that (a) the papers are inherently similar and are difficult to separate via k-means, (b) the parsing routine requires a longer stop word list, (c) different facets of the topic are gradually emerging at each new cluster size, (d) each paper opens with an introduction which covers similar ground to every other paper in the topic and (e) constant references to prior work, especially highly cited work, add similarity.

To evaluate the quality of the clusters found, the formula for incomplete normalisation (see Appendix D) was used to estimate the size of the similarity between two documents of which could be considered significant. The sizes of the VSM representations in the Dark Matter corpus were ranked and the 10^{th} percentile was chosen as a

representative "small" document. There were five documents at this size of 269 terms. For all of these five documents, the median term frequency was equal to 1 and the third quartile term frequency was equal to 2, a reasonable expectation given the power law distribution of term frequencies described by Zipf's law. Choosing a reasonably frequent term $f_n = 4, g = 2$ to maximise its effect and using the term frequencies, f_i, of the five selected small documents ($n = 269$) to calculate the term $\sum_{i=1}^{n} f_i^2$, the error estimate

$$\text{error estimate} = \frac{g^2 \sum_{i=1}^{n} f_i^2}{(\sum_{i=1}^{n} f_i^2 - f_n g)^2} = \frac{4 \sum_{i=1}^{269} f_i^2}{(\sum_{i=1}^{269} f_i^2 - 4 \cdot 2)^2} \tag{4.9}$$

was calculated for each document. The average error estimate over the five documents was found to be 0.001. This value has been used throughout the comparisons of similarity within and between documents in the Dark Matter corpus to establish a threshold of significance. Any difference in similarity below this threshold of 0.001 can be reasonably set aside as being "small".

Using the best text clusters found at $k = 4$ (the clusters with the lowest $\widehat{RSS}_{min}$ value), the centroid of each cluster was re-calculated and the similarity of each document with each cluster centroid was found. Seven documents were found to have a difference in the similarities between the two closest centroids of less than the error estimate. No strong correlation between these documents seems to exist in either their cluster membership or their topic, as inferred from their title. This set of clusters has a fairly even distribution of papers, the four clusters containing 459, 527, 611 and 763 papers respectively.

In the debate between hard and soft clustering, that only 7 documents out of 2659 are sitting near a boundary provides a good cause for being satisfied with hard clustering, depending on the characteristics being sought from the data. In this chapter, only an indication that papers could be separated into distinct text clusters (regardless of sharing characteristics with other clusters) and whether a natural number of clusters could be deduced form the clustering process.

4.8 Cross-validating Louvain Clustering with k-means

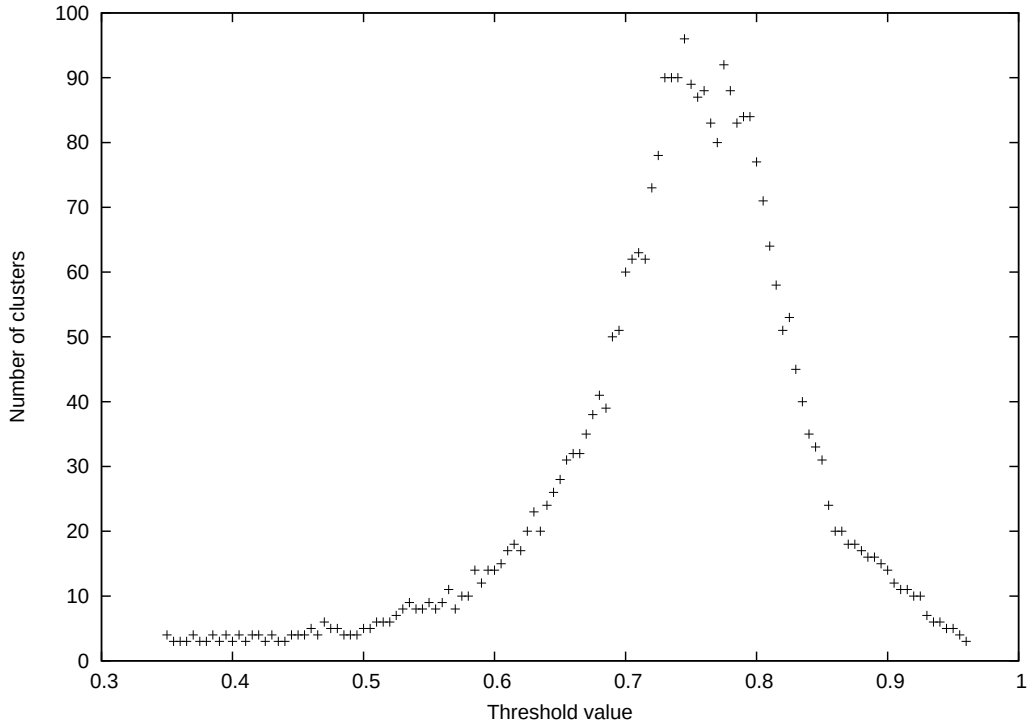

Figure 4.6: Number of text clusters, K, found using the Louvain algorithm on a document similarity network produced by adding a link between documents that have a text cosine similarity greater than a threshold value, θ.

Transformation of a problem can make other tools available for analysis. If the document similarity matrix is reframed as a graph then document clusters can be found using network community finding algorithms used on graphs, such as the Louvain algorithm used in Chapter 3. In a matrix of text similarity, the elements are the similarity between the pair of documents represented by the row and column indices. As the cosine similarity has been calculated between all pairs of documents (presented in Section 4.7), all elements of this similarity matrix are in the range $[0, 1]$. To transform the matrix into a graph, consider each document as a node with the link between

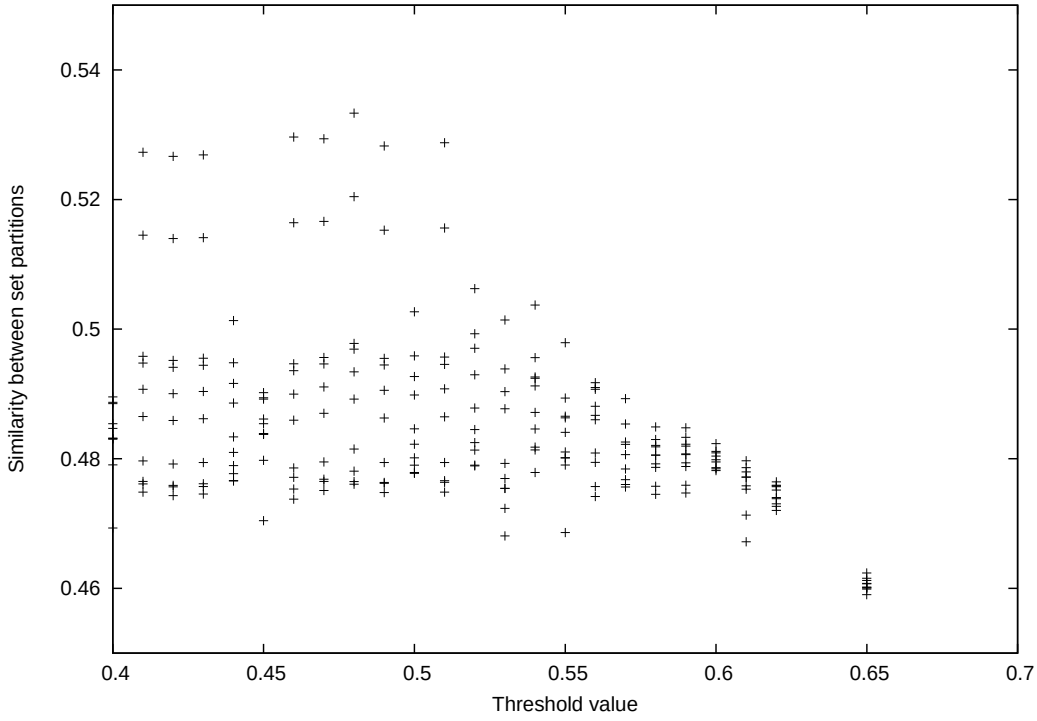

Figure 4.7: Rand Index comparing text clusters found using Louvain on a thresholded document similarity network (Figure 4.6) with those found using k-means clustering (Figure 4.4) with the same number of clusters.

documents i and j, given by a_{ij}, connecting the two documents if the cosine similarity between them, σ_{ij}, meets or exceeds a chosen threshold, θ, such that

$$a_{ij} = \begin{cases} 1, & \sigma_{ij} \geq \theta \\ 0, & \sigma_{ij} < \theta \end{cases}$$

This representation produces what in network terminology is known as an adjacency matrix. The similarity matrix is symmetrical ($a_{ij} = a_{ji}$) and is represented as an undirected network. Because every document is identical to itself, the diagonal elements are set to zero ($a_{ii} = 0$) to remove self-loops from the network which may adversely affect the clustering algorithms. Documents which become disconnected from the network by virtue of having no text similarity values greater than or equal to the threshold are removed from the network.

Choice of the threshold is arbitrary. To understand the behaviour of this choice, the maximum number of clusters found using the Louvain algorithm was plotted for threshold values running from 0.35 to 0.96 in steps of 0.005. The result is a curve which peaks in the vicinity of 0.755 ± 0.02. As the threshold increases, the number of links decreases and from 0.8 upwards the number of nodes in the graph also decrease because there are fewer documents sufficiently similar with which to link. To compare the two methods, the Rand Index for measuring the amount of agreement between two partitions was computed using the Perl module `Set::Similarity`. This algorithm requires that the two sets are identical and that both sets contain the same elements. Because the thresholding method can disconnect some documents from the network, an extra subset in the thresholding partition was created from all the missing elements, this being the union of both sets. Computing the Rand Index between thresholded and k-means clustered partitions found spread around 0.5. As the Rand Index can be described as the number of agreements between the two sets divided by the number of agreements *and* disagreements, it was felt that the use of an extra set could be reducing the measurement to 50% accuracy. These are shown in Figure 4.6.

To check what similarities could be expected if the methods were equivalent, partitions were compared with other partitions created with the same method. The results, shown in Figures 4.8 and 4.9, demonstrate that the methods are self-consistent, producing high similarities even for large discrepancies in number of clusters between partitions. A linear fit to the data was plotted to show the average trend of the data in spite of the crowded plots.

The possibility of including similarities between partitions with different numbers of clusters was explored in Figure 4.10. The previous calculations were repeated including k-means partitions that had up to 5 clusters more or less than the thresholded partition. These similarities were grouped by difference in number of clusters, ΔK, and fitted with weighted cubic splines to show their trends. To reduce visual clutter, only the data for $\Delta K = 0$ and $\Delta K = 5$ were plotted. It is possible to see that the fits for $\Delta K = 0$ and $\Delta K = 1$ are almost identical, whereas the fits for higher ΔK surprisingly have higher similarities. The reason for this discrepancy has not been investigated.

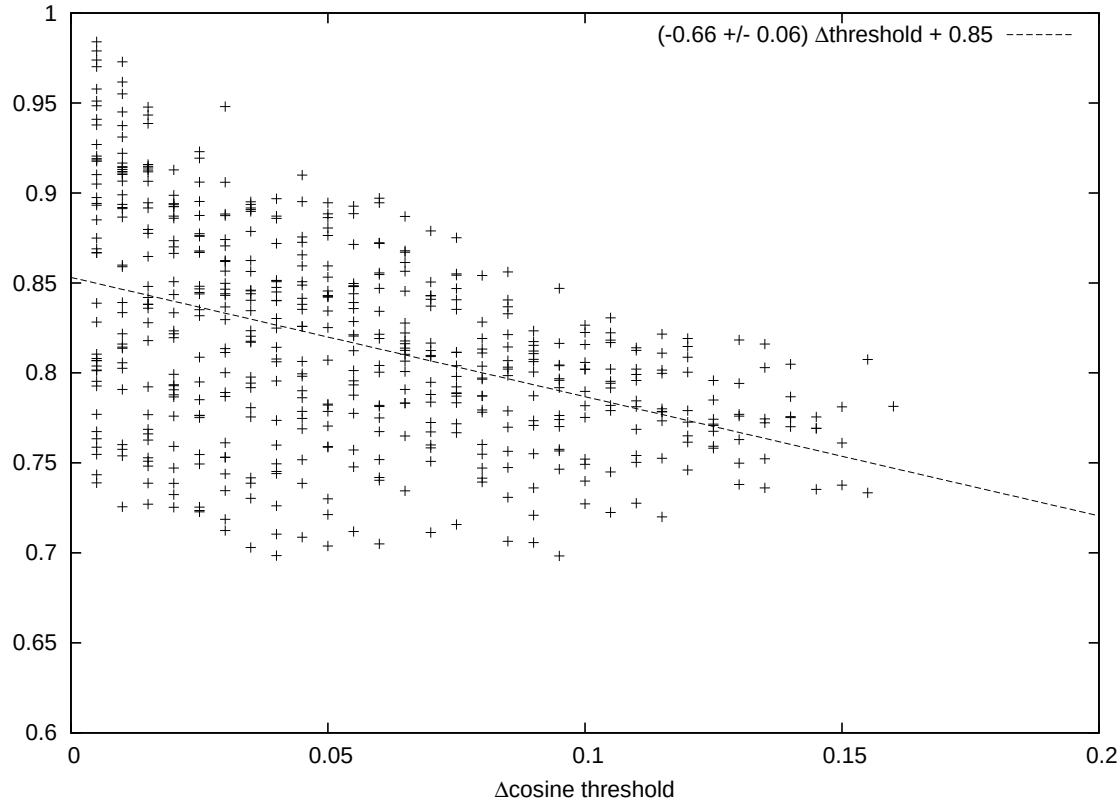

Figure 4.8: Rand Index set similarity of two partitions of Louvain communities found on thresholded document networks plotted against the difference in the thresholds used to produce the document networks. A linear fit to the data is plotted for comparison of the trend.

To determine if this is an artefact of the increase in number of clusters, the same plot was produced by shuffling the elements of the k-means clusters (a technique sometimes known as a Monte Carlo simulation) when measuring the set similarity with the Louvain partitions. It is clear from Figure 4.11 that indeed the intersection of the sets performs better than random, but the upward trend and the comparison's improvement is a feature of the measurement technique on higher number of clusters suggesting that lower values of K are more powerful indicators of similarity. The strength of this effect above randomness was plotted in Figure 4.12 from the differences between the values from the Monte Carlo simulation and the thresholded data values,

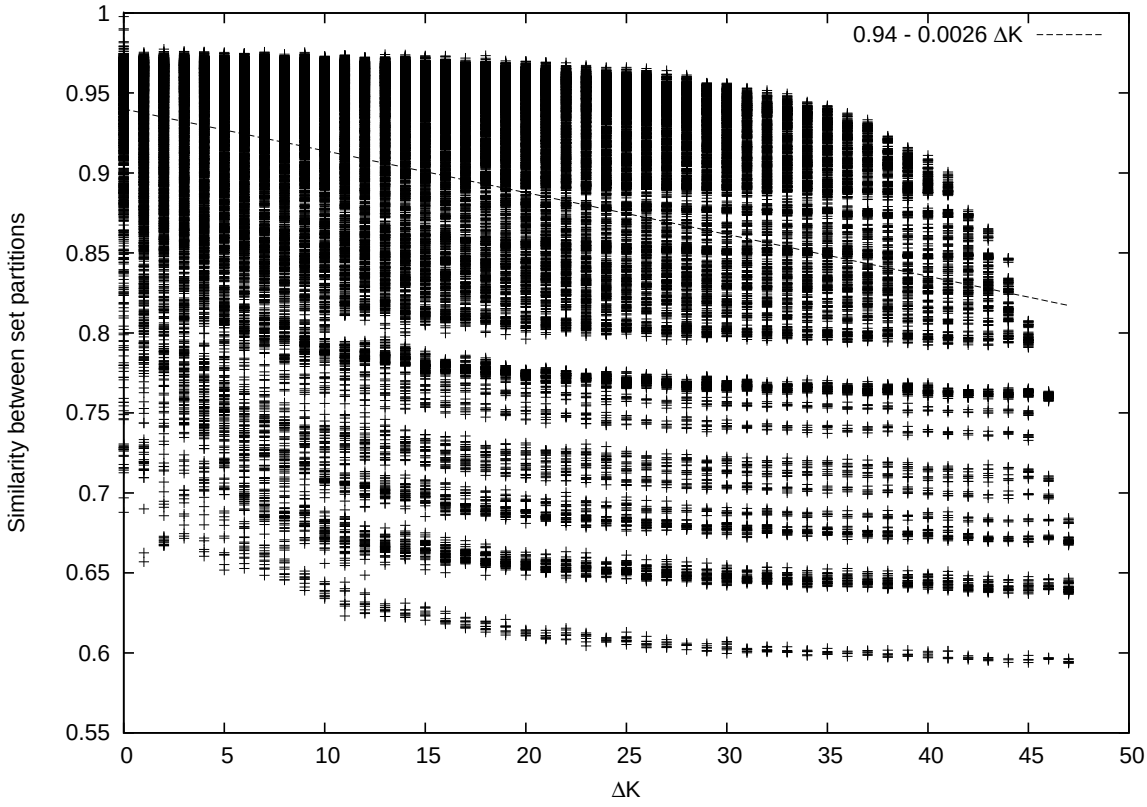

Figure 4.9: Rand Index set similarity of two partitions of k-means clusters plotted against the difference in the number of clusters found, K, for each partition. Because of the large number of data points overlapping on the plot, a linear fit to the data is plotted for comparison of the trend showing that most of the data lies near the top of the plot.

then dividing by the simulated values to provide a ratio.

The result of these investigations is that similarity automatically improves for higher K. Researchers should be aware of this behaviour in their evaluations. The similarity of different numbers of clusters are acceptably close to similarity measurements of equal number of clusters which may be of use in situations where more values are required. The actual effect found from similarity measurements is more pronounced at low K. Louvain clustering on thresholded similarity matrices and k-means unsupervised learning methods produce similar but not equivalent results. The advantage of

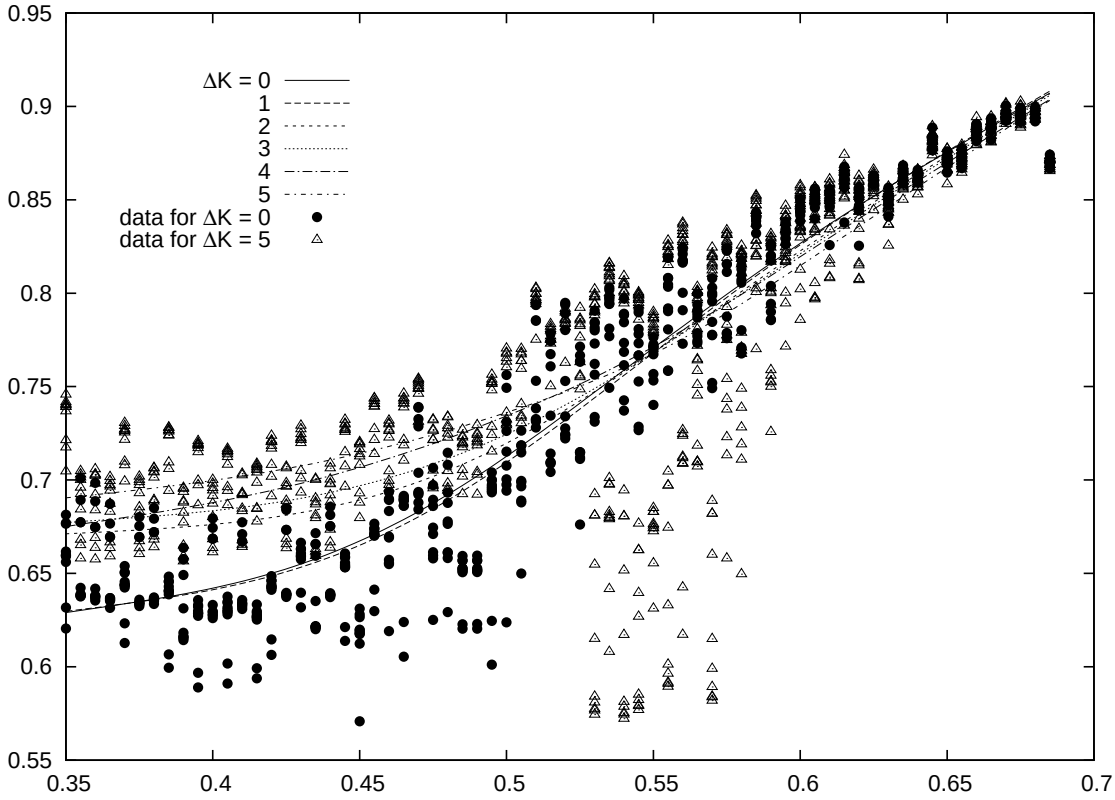

Figure 4.10: Rand Index set similarity (y-axis) between set partitions for a document network of threshold, θ, (x-axis) and best text clusters found with k-means where ΔK is the difference between the number of text clusters and the number of Louvain communities found on the document network. Splines are fitted for all $\Delta K = [1, 2, 3, 4, 5]$, but data points are only plotted for $\Delta K = [1, 5]$.

the Louvain clustering technique is the speed with which it provides answers as long as it has the entire corpus to process, whereas the side effect of k-means clustering is the production median vectors for each cluster which can be used to label and represent the cluster in classifying new documents presented after the clustering process.

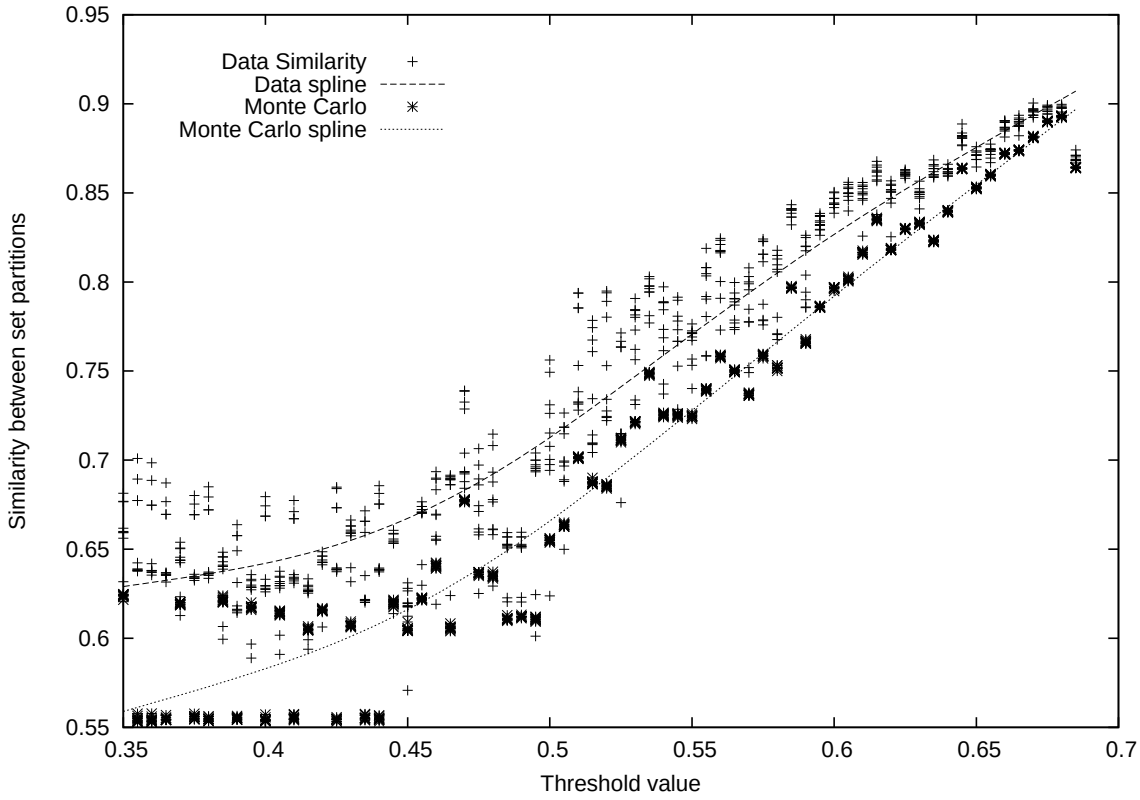

Figure 4.11: Monte Carlo simulation in comparison with the Rand Index of set similarity between set partitions found from Louvain communities found for a document similarity network produced with a threshold value, θ, for θ from 0.35 to 0.69. Splines were fit to the data to show the average trend of the data.

4.9 Cluster Labelling

The result of unsupervised learning is the ordering of information. The assignment of labels gives greater insight into those clusters. Although a number of people have reviewed many different clustering techniques [Steinbach et al., 2000, Berkhin et al., 2006] prior to Manning et al. [2008], few have had much to say about the labelling process. Weiss et al. [2010a] says that creating good labels can be expensive, referring to the time taken by an expert to assess the usually large document collection and derive meaningful labels. The ADS has chosen to invest that effort in developing an

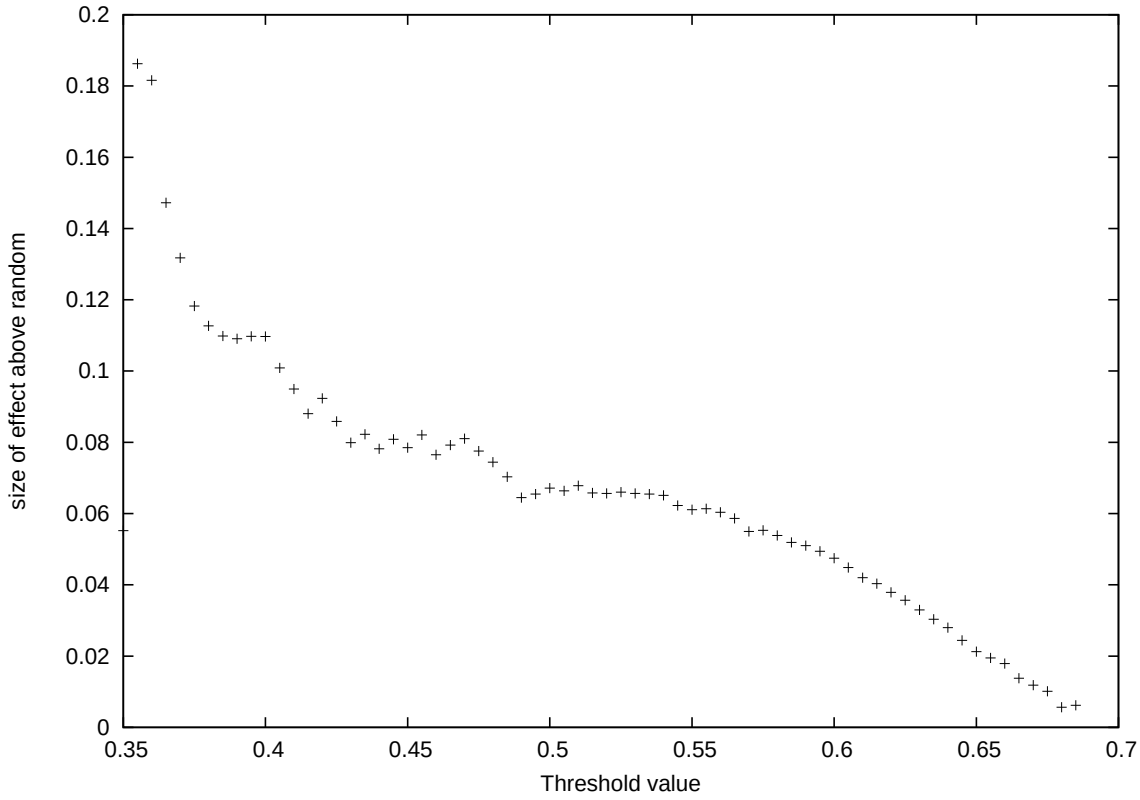

Figure 4.12: Magnitude of the effect in Figure 4.11 — Rand Index set similarity of Louvain clusters with k-means compared to a Monte Carlo simulation of the k-means partitions, as a ratio of ΔSimilarity/Similarity.

ontology [Accomazzi and Dave, 2011]. Producing categories requires *a priori* knowledge about most, if not all, of the collection. A more satisfying solution is to extract the labels from the clusters themselves. Stein has described a framework for deriving label information that they designate "weighted centroid covering" [Stein and Eissen, 2004]. Cattuto et al. [2007] outlines efforts to reduce the individual workload by engaging a community in collaborative tagging.

With an increase in interest in labelling clusters and knowing that the interpretation of clusters is cognitively demanding and not well supported, Chen et al. [2010] developed tools to assist the data miner which are incorporated into the CiteSpace tool

[Chen, 2006]. They delve deeper into NLP using noun phrases as labels and evaluating nine different ranked lists constructed from the keywords and noun phrases extracted from the titles and abstracts. By relying on noun phrases (the extraction process was not described), they may be ignoring the effects of synonymy. They find that noun phrases from the titles ranked by *tf-idf* is an effective method for labelling, but it is dependant on the quality of the available terms. One source of quality terms would be to use extracted topics if a topic model of the subject is available.

Hulpus et al. [2013] uses graph-based methods on text to identify topics assisted by the external knowledge base, DBpedia, to provide word-sense disambiguation between concepts [Lehmann et al., 2015]. They highlight the many challenges facing labelling with fully automated systems not being realised. One recent attempt at automated labelling uses the clusters from unsupervised learning to train an Artificial Neural Network (a type of supervised learning) and deriving labels from the attributes activated during the testing phase of the neural network [Lopes et al., 2016]. It is an intriguing approach, but its use of a relatively small number of member attributes (as appropriate in numerical Data Mining) to train the ANN presents a challenge to adopting the method in the high-dimensional realm of Text Mining.

These endeavours are not surprising. The document collection was large enough to warrant automated clustering. It follows that automated labelling would be preferred, if it were easy. One method is to simply take the most frequent terms in the centroid, which are calculated during the k-means processing, as the cluster's label, yet the meaning of this label is uncertain. Just as the VSM loses semantic information when word order is discarded, it is neither straightforward or unambiguous to construct meaning from a list of words. Weiss suggests that weighting the terms in the label by their utility to discriminate between documents may improve the quality of the labels, in the same way that *tf-idf* was used to produce more contrasting document vectors. Another method is to simply use the title of the document closest to the centroid as it is assumed to be the document most representative of the cluster.

Attempts at labelling the clusters found using k-means demonstrated the difficulty of obtaining clear, meaningful labels automatically. Each cluster was labelled with

the 10 most frequent keywords assigned to those papers in the ADS after the four most common keywords (dark matter, galaxies, astrophysics, cosmology) has been removed.

It was noted that while high energy physics was present somewhere at all values of k, MOND was not encountered until $k = 31$. When the number of keywords per label was doubled, MOND still only appears at $k = 24$. Although it is a radical approach to the Dark Matter problem, its lack of prominence in keywords could be attributed to its need to discuss the theory in comparison to the standard theory, whereas mainstream discussion takes place with little or no mention of MOND.

4.10 Summary

This chapter has shown how Text Mining differs from Data Mining and how it relates to the fields of Artificial Intelligence, Natural Language Processing and Information Retrieval. The extraction of meaning from text was discussed within the context of Linguistics and NLP. The connection was then followed from NLP to Artificial Intelligence where the use of ontologies to effect knowledge representation can be a key component in machine reasoning. Specifically the OWL ontology describes all possible citation events. Supervised and unsupervised learning, two important foci of AI, when applied to Text Mining are Classification and Clustering.

In order to cluster, a method for comparing documents is required. The Vector Space Model was introduced as a tractable representation that has an accessible metric of similarity using the definition of the dot product of two vectors. This was justified on the basis that it retains 80% of the information. The LaTeXML project was invaluable for the creation of document vectors from primarily LaTeX sources. The standard processing from Information Retrieval was then described and included an upper bound on the error due to incomplete Normalisation. Stemming and Stopping were discussed and their implementation described.

The distribution of similarities was explored through different document sections and how text changes over time. Two treatments for calculating the similarity, cosine

distance with k-means and *tf-idf* using a threshold value on the similarity matrix, were described and evaluated by comparing the clusters they both produced using the Rand index. These clusters were labelled using their most frequent keywords, but were not found to have the clarity of human-applied labels.

This chapter demonstrates how meaning can be extracted computationally from the published research in an area of science and how its social dynamics may be studied through how authors explain their ideas and findings. The papers they produce can clustered by topic and even treated as graphs. The methods used to compare clusters will be used again in the next chapter.

5 Observations on Communities

While Complex Networks and Data Mining are mature fields attracting copious amounts of serious study, the combination of techniques from both areas is a rare occurrence. One reason for this lack is the depth of knowledge required for entry into each area. Regardless, the space between two active fields is fertile ground for exploration using well-honed tools in a new situation, provided that commonalities exist between the two. An academic paper is obviously a document with which the text mining community is very familiar and the citations between papers have been studied many times before as a complex network. This chapter is a study in how cross-over approaches fare with these fields.

5.1 Categorisation of the Dark Matter corpus

As noted in Section 3.3, communities can be defined as groups with more links within the group than outside of them, but the nature of the communities formed required some deliberation. Papers deposited in arXiv are assigned a category which is useful for separating some of the papers based on topic. The fact that papers can overlap two or more topics is revealed in more recent papers when arXiv began to allow multiple category assignments. While High-Energy Physics is conveniently divided into experiment, phenomenology and theory, no such arXiv category divides observational astronomy from theoretical astrophysics with the exception of cosmology in `gr-qc`.

The papers constituting the Dark Matter corpus are grouped according to their arXiv category in Table 5.1 with the majority in astrophysics. The next largest category is High Energy Physics with just over 10% of the papers.

The ADS search engine was used to identify which papers in the corpus correspond to work on MOND. Of the 102 results, 95 correspond to `astro-ph`, 6 to `gr-qc` and 1 to `physics.gen-ph`. A subset of 21 papers in the MOND section of `astro-ph` also had subcategories: 47% in `gr-qc`, 30% in `hep-ph` and 23% in `hep-th`.

category	*description*	*number in corpus*
astro-ph	Astrophysics	2228
cond-mat.other	Other Condensed Matter	1
gr-qc	General Relativity and Quantum Cosmology	94
hep-ex	High Energy Physics – Experiment	39
hep-ph	High Energy Physics – Phenomenology	207
hep-th	High Energy Physics – Theory	32
math-ph	Mathematical Physics	1
nucl-ex	Nuclear Experiment	6
nucl-th	Nuclear Theory	3
physics.ao-ph	Atmospheric and Oceanic Physics	1
physics.comp-ph	Computational Physics	1
physics.flu-dyn	Fluid Dynamics	1
physics.gen-ph	General Physics	34
physics.ins-det	Instrumentation and Detectors	12
quant-ph	Quantum Physics	1
	total	2661

Table 5.1: Categories of Papers in the Dark Matter corpus

The subcategories show interesting features as the least common examples lead the way to interesting segues on the topic, including the involvement of "volcanogenic dark matter" in extinction events in Earth's geologic history which was filed under `physics.bio-ph` and `physics.geo-ph`.

5.2 k-core visualisation

Networks of any significant size pose difficulties in their visualisation due to the cluster produced by criss-crossing links between nodes with little space available for labels. Node separation is an arbitrary choice but necessary in order to resolve details within the network. Various algorithms for graph layout attach meaning to the distance between nodes allowing the researcher to intuit understanding about the network from its structure. k-core produces a low-complexity visual as an analytic tool for examining

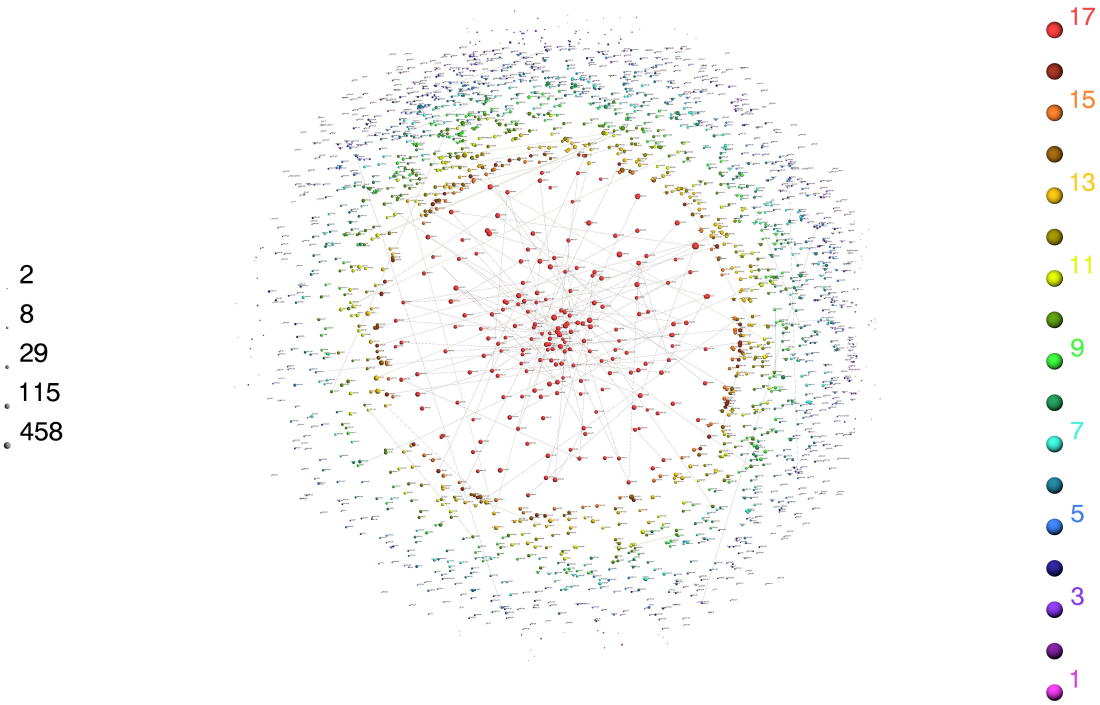

Figure 5.1: k-core visualisation of the Citation Network core papers. The grey legend on the left represents the degree of the node, while the coloured legend on the right represents the k-shell index.

the internal organisation of the network structure. Described in Alvarez-Hamelin et al. [2005] and Beiró et al. [2008], k-core graphs were produced using `LaNet-vi` for the core citation network and the k-means similarity graph to permit direct comparison. The size of each node represents its degree. Hubs, therefore, are the larger spheres. The colour scale separates the nodes according to their k-core shell index, where the index or "coreness" is the connected maximally induced subgraph of all nodes that have a degree of at least k. The diameter of each shell is related to the number of nodes in the shell, with nodes moved closer to nodes with which they share more links.

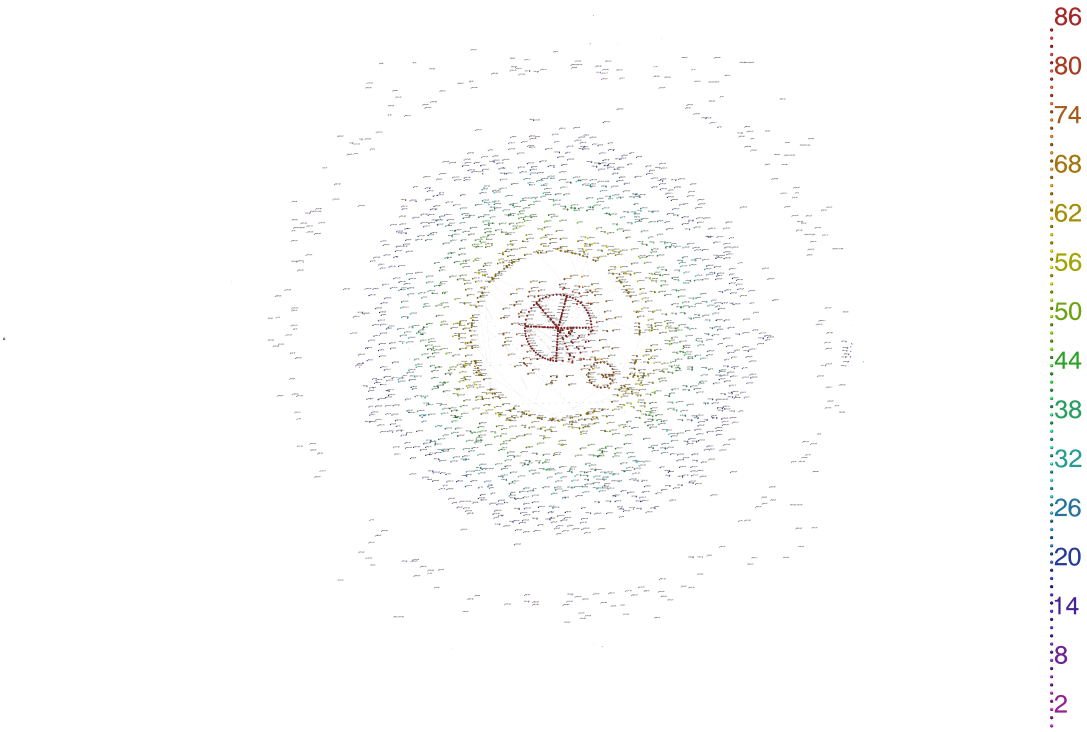

Figure 5.2: k-core visualisation of the document network produced from the VSM model where documents sharing a similarity of 0.5 or more are linked in the network as described in Section 4.8

The visualisation of the citation network, shown in Figure 5.1, identifies the range of node sizes, seen previously in Figure 3.8, and shows that degree is strongly correlated with k-core shell index: hubs are closer to the centre of the figure and there are no star-like subgraphs. Since a node is moved closer in the diagram to the nodes to which it is connected, thick shells show that this network has a wide range of nodes connected to different higher shells. This is an indication of a disassortative network, such as examples of Internet Routing networks. Section 3.6 briefly discusses the implications. A closer examination of the labels reveals that papers on MOND are found in the lower

half of the figure while papers on high-energy physics are spread through the upper half with the centre dominated by papers in the `astro-ph` category.

In contrast, the similarity graph produced by linking all Dark Matter core papers with a text similarity of 0.5 or more, shown in Figure 5.2, has little or no fluctuation in degree, with clearer distinctions between shells and paper categories evenly distributed over the diagram. The circular segmented structures in the centre of this figure are separate entities of related papers, the smaller one hosting a number of high-energy physics papers which is disconnected from the main similarity graph. The disparities between these two visualisations suggest that a simple mapping from one domain to another is unlikely to exist, a judgement that would be difficult to arrive at with confidence through statistics alone.

5.3 Correlations

For each pair of papers, the cosine similarity between the two was computed and the value averaged with all the other pairs belonging to the same two groups found by the Louvain algorithm. When a pair both belong to the same group, they are said to be within the community. When the pair are from different groups, they are outside the community or external to it. The average similarity among groups is presented as a heatmap in Figure 5.3 to highlight the differences between and within groups of over 450 values. This type of representation quickly identifies the relationships between the groups which are laid out in Tables F.3 and F.4. The colour scale uses blue for less similar (minimum value 0.09) and red for more similar (maximum value 0.77) with white set at a value of 0.26, the mean of the average similarities among the groups. Light colours represent groups of near-average similarity with darker colours being of interest. Black is used for missing values. To present similar groups together, they are ordered along the axis by similarity to their neighbours. Only groups with at least two core papers are presented in order to judge the characteristics of a group rather than an individual paper. They are labelled with an ad-hoc numeric identifier.

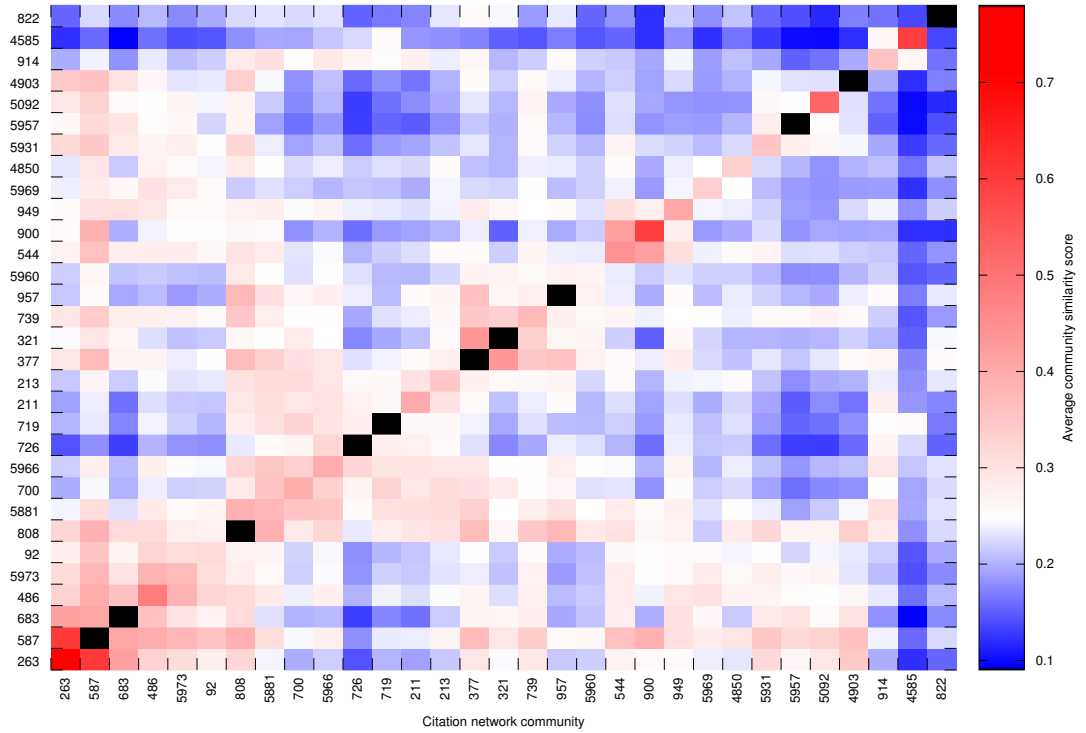

Figure 5.3: Correlations of average text similarity between and within citation network communities from the 31 communities containing at least 2 papers from the Dark Matter corpus.

It can be seen that along the diagonal axis from lower left to upper right that the papers within a group exhibit stronger similarity, whereas off-axis, there are few departures from average similarity. This is confirmed by calculating the similarity of all papers to papers outside their group to be 0.25 ± 0.10 and within their group, the similarity is 0.34 ± 0.12 .

With a column of many shades of strictly blue, **Group 4585** is markedly different from all other groups. The 8 core papers that are included in this group are all from the majority arXiv category `astro-ph`. The keywords galaxy, white dwarfs, halo, luminosity

function are not particularly illuminating unless the involvement of white dwarfs in Dark Matter is a self-contained sub-topic progressing independently from the rest of the field. Similarly, **Group 822** has only 2 papers with the keywords cosmology and dark matter The reason for the distinctiveness of these two groups does not become apparent until further exploration which is described in Section 5.6.

The two off-axis red squares indicate a strong textual similarity between Groups 263 and 587. Both groups, while in the `astro-ph` category, focus on the phenomenology of high-energy physics. Although their average similarity is high, 0.605 ± 0.165, and the link density is low (3 out of a maximum 6), it is not advisable to draw strong conclusions from these values as the numbers are based on three papers from **Group 263** and two papers from **Group 587**. After investigating the five papers, it was found that one of the two papers from 587 had referenced all three of the papers in 263 in comparison to six references that **Group 263** has amongst its own members.

While making these remarks about similarity between groups, it is useful to consider them in the context of the distribution between individual pairs, shown in Figure 4.1, which is broad and skewed to lower values with a peak near 0.25 in textual similarity. Further to explore the difference in textual similarity between the groups of citation network communities, the distribution of similarity values of papers within a community and between different communities was plotted in Figure 5.4 with the average similarity value for both distributions indicated. The large number of values available—373 294 pairs of papers within a community and 2 386 781 between different communities—allow for smooth plots for the distribution and excellent statistics. Student's t-test is a statistic for measuring the significance of the difference of two mean values. The null hypothesis is that the two means are drawn from the same distribution. Selecting a 99% confidence level, the calculated statistic of $t(452479) = 458.9$ rejects the null hypothesis. The distribution plots in Figure 5.4 overlap and share the same skewed shape, but are distinctly different distributions, confirming what was found with the Student's t-test. On average, the textual similarity of a paper to papers within the same citation network community is measurably higher than with papers outside of that community.

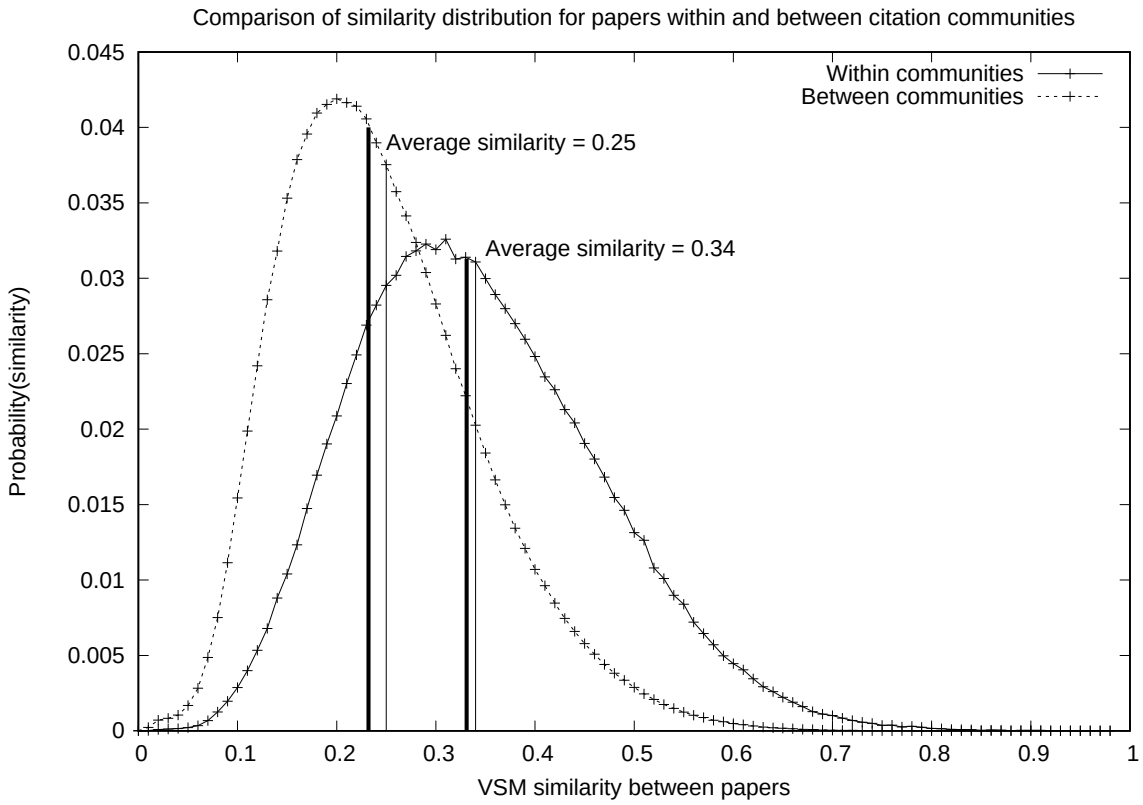

Figure 5.4: Correlations of similarities between and within citation network communities. The average VSM similarity value for each distribution is marked with a thin black line. The median similarity is indicated with a thick grey line. The probability values plotted on the Y-axis are number of paper-pair similarities in a given bin (width 0.01 on the X-axis) divided by the total number of pairs in the distribution.

5.4 Adjusted Rand Index

Revisiting the technique of measuring the similarity of partitions from Chapter 4, the Adjusted Rand Index was calculated for all four levels of clusters found in the citation network via the Louvain algorithm against the k-means text clusters with the lowest RSS values for each $k \in [3..50]$. These were then plotted together in Figure 5.5 where it exhibits the k-means clusters that use `Lingua::EN::StopWords` while the text clusters in Figure 5.6 also removed the stopwords in Table F.2. Features that

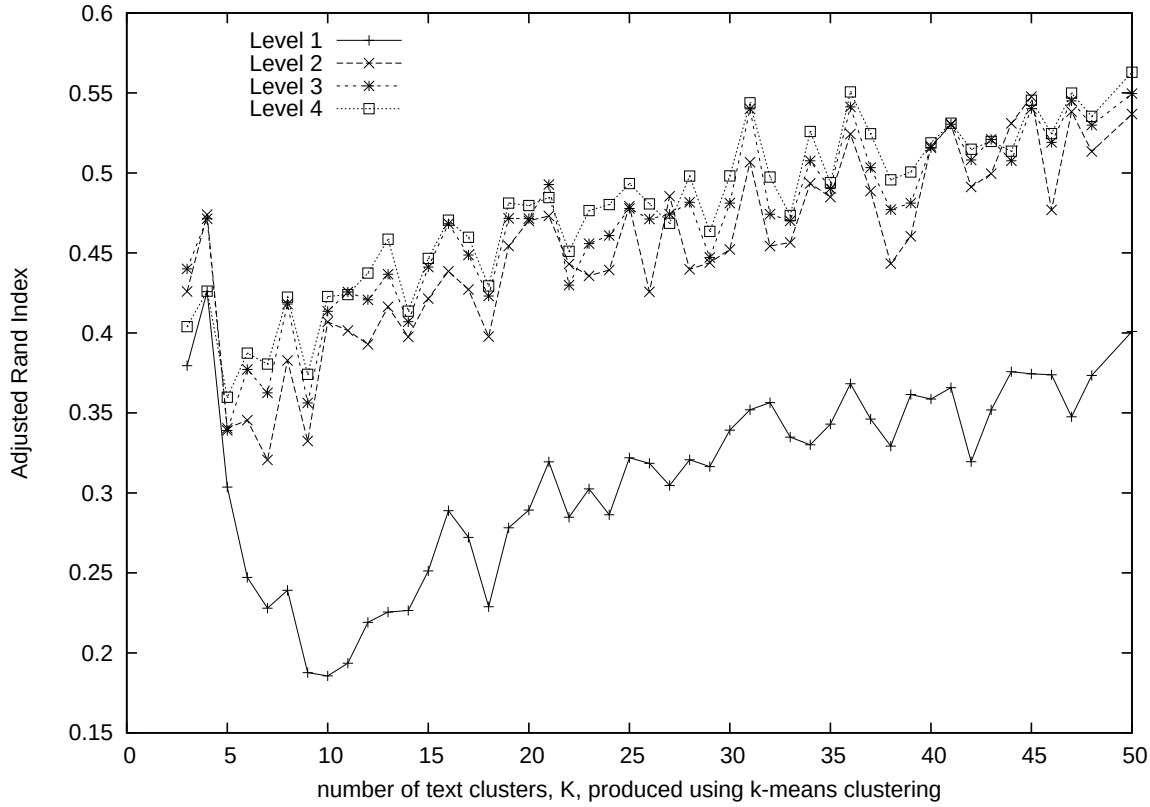

Figure 5.5: Adjusted Rand Index of partitions found from citation network communities (Levels 1–4 of the Louvain algorithm) against K text clusters (stopped using the Lingua::EN::Stopwords list and clustered with k-means).

are common between the two figures are the peaking of all four levels of the network hierarchy followed by a sharp drop and gradual, if erratic, rise as k increases. Also, the lowest level of the hierarchy with the greatest number of groups, level 1, suffers a much greater drop than any of the other levels. Other features are the peaks and dips coinciding in levels 2–4, with peaks in (a) at $k = 32, 36$ and (b) at $k = 10$ while dips are found at $k = 9, 15, 18, 37$ in (a) and $k = 7, 12, 24, 35$ in (b). These other features are not shared between the two sets of k-means clusters, leading to an assumption that they are artefacts of the unsupervised learning process, not indicative of the underlying structure.

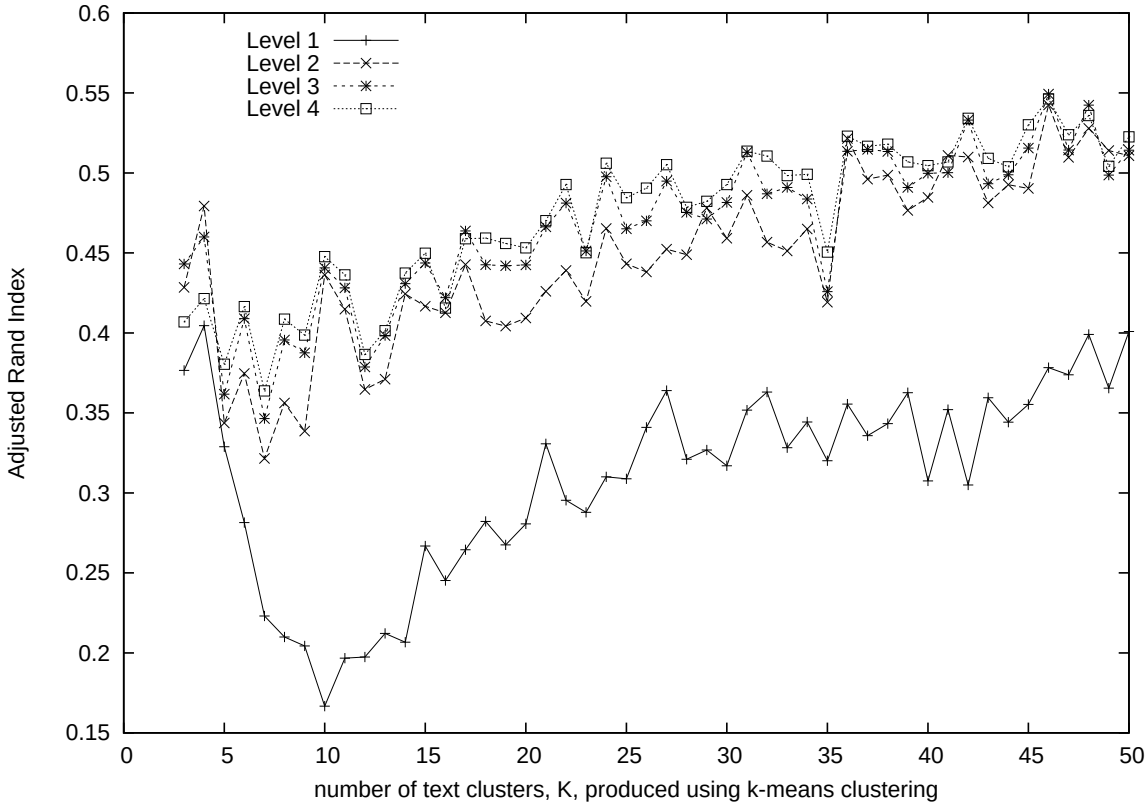

Figure 5.6: Adjusted Rand Index of partitions found from citation network communities (Levels 1–4 of the Louvain algorithm) against K text clusters (stopped using the Lingua::EN::Stopwords list plus extra stopwords chosen from *tf-idf* ranking of terms from the corpus).

5.5 Vector Space Model as a network

Further exploring the structure of the k-means clusters, the document graph was created by linking papers in the Dark Matter corpus with a similarity greater than or equal to a threshold value, θ. This approach allows the use of concepts and tools well-honed in the study of complex networks to attack text mining problems. The degree distributions for the document graphs with $\theta \in [0.3, 0.4, 0.5, 0.6]$ were plotted in Figure 5.7. At the lowest threshold of 0.3, the distribution is a broad dome with

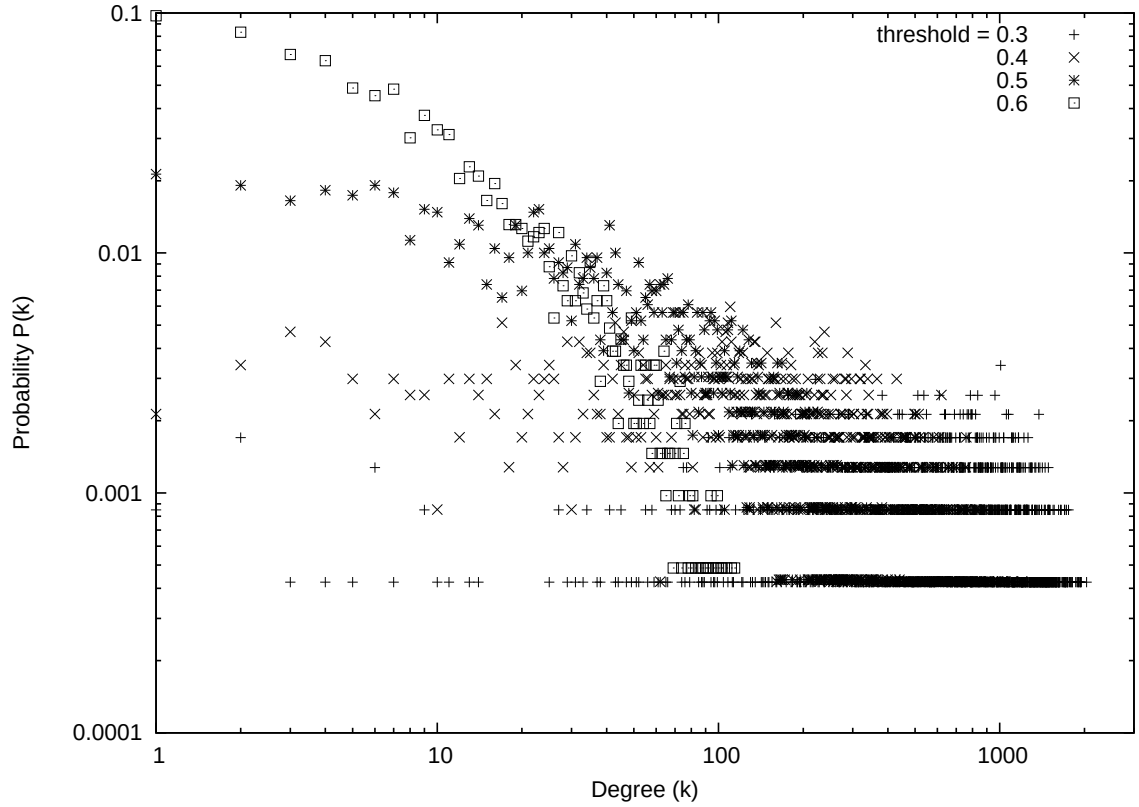

Figure 5.7: Degree distributions for document graphs created with thresholds, $\theta \in [0.3, 0.4, 0.5, 0.6]$, examining how the degree distribution changes with θ and which document graphs have power law degree distributions.

proportionally very many highly-connected papers. At the next lowest threshold, 0.4, the distribution becomes skewed toward increased numbers of lesser-connected papers. Only at the threshold of 0.5 does P(k) begin to develop the long tail and at 0.6 (shown in Figure 5.8), it is somewhat like the power law distribution characteristic of many complex networks. The increasing threshold, listed in Table 5.2, sees a dramatic drop in the number of links, while the number of nodes gradually decreases until a larger drop at 0.6.

In relation to the formal definition given in Section 3.1.3, all text graphs show the small world property with $\frac{\langle l \rangle}{\langle l_r \rangle} < 2$, however, only the graphs at threshold $\theta = 0.5$

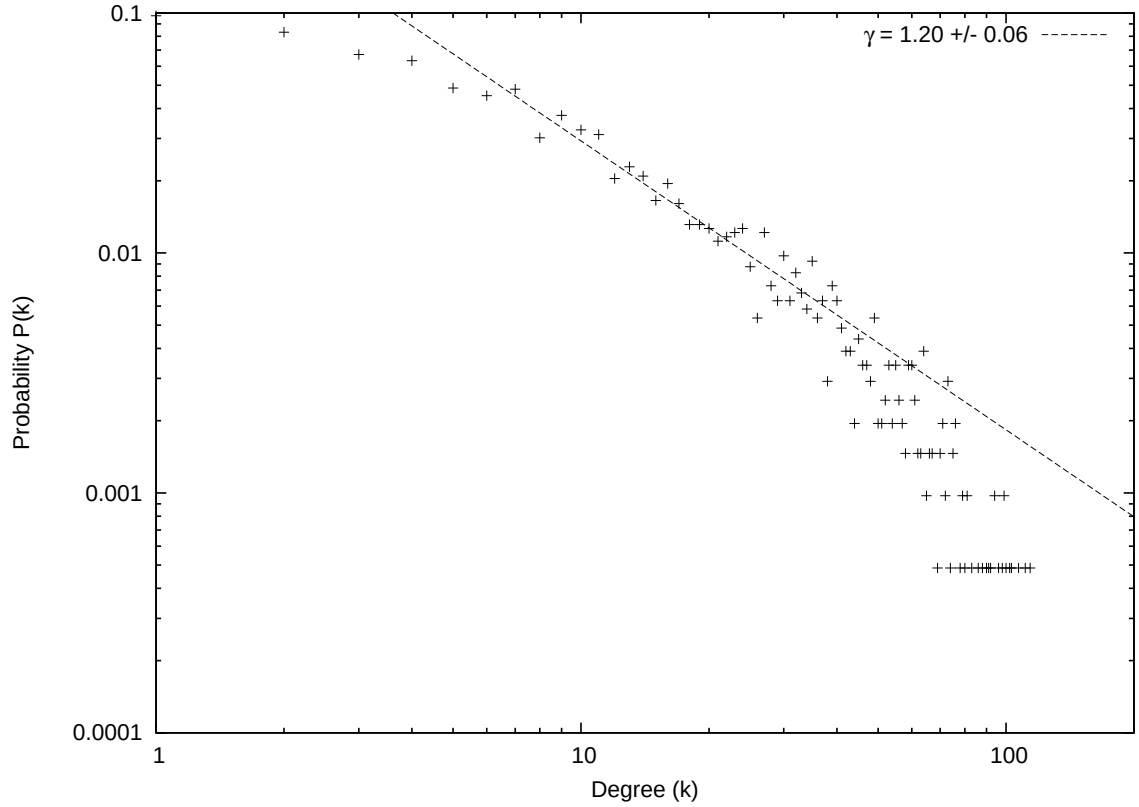

Figure 5.8: Degree distribution of document graph with threshold, θ = 0.6 fitted with $P(k) \propto k^{-\gamma}$

and above have significant clustering, $\frac{C}{C_r} > 10$, and the long tail of a scale-free degree distribution (as seen in Figures 5.7 and 5.8) to qualify as complex networks according to the formal definition.

5.6 Community Identity

Figure 5.9 was produced using the `Force Atlas` algorithm in the Gephi software package for networks. Due to computational constraints, the network visualised consisted only of the core and first shell of the Dark Matter network (47 187 nodes, 161 870

θ	N	L	$\langle k \rangle$	$\bar{C}$	$\bar{C}_r$	$\bar{C}/\bar{C}_r$	$\langle l \rangle$	$\langle l_r \rangle$	$\langle l \rangle / \langle l_r \rangle$
0.3	2353	858 339	730	0.70	0.31	2.3	1.71	1.18	1.4
0.4	2345	299 457	255	0.62	0.11	5.6	2.10	1.40	1.5
0.5	2300	81 176	70.6	0.57	0.031	18	2.83	1.82	1.6
0.6	2055	16 228	15.8	0.54	0.0077	70	4.41	2.76	1.6
0.7	1328	2 505	3.77	0.55	0.0028	196	8.51	5.42	1.6

Table 5.2: General network statistics for the document graph formed by adding links between papers with similarity $\geq \theta$. As the threshold θ increases, the number N of connected nodes in the document graph decreases, as do the number of links, L and the clustering coefficient, $\bar{C}$. The mean path length $\langle l \rangle$ increases. The clustering coefficient for the document graph is much greater than for a random graph, $\bar{C}_r$, of the same size.

links). The nodes were coloured according to the largest seven communities found by the Louvain algorithm with, in decreasing order of size, purple for traditional astrophysics (**Group 213**), green for high-energy physics (**Group 92**), blue for cosmology and gravitational lensing (**Group 211**), red for Dark Energy and alternate theories (**Group 739**), black for detection of particles and high-energy physics (**Group 5973**), orange for Dark Matter models and galaxy clusters (**Group 914**), and turquoise for MOND (**Group 691**). The labels for these communities were arrived at by examination of the titles of the papers in the communities.

The links are not shown to reduce visual clutter. The Force Atlas algorithm is a layout using physics equations for dynamics that considers all nodes to be protons (repulsive) and links to be springs (attractive), meaning that all nodes spread out as much as possible from each other while being pulled towards nodes to which they are linked.

This visualisation shows that some communities overlap, sharing many links, such as the green and black clusters of high-energy physics and particle physics. The fringe community, Dark Energy, connects with Dark Matter via models (orange) and MOND (turquoise). Concerned with using observations to evaluate models for both Dark Matter and its alternatives, MOND is a long, thin community that touches many

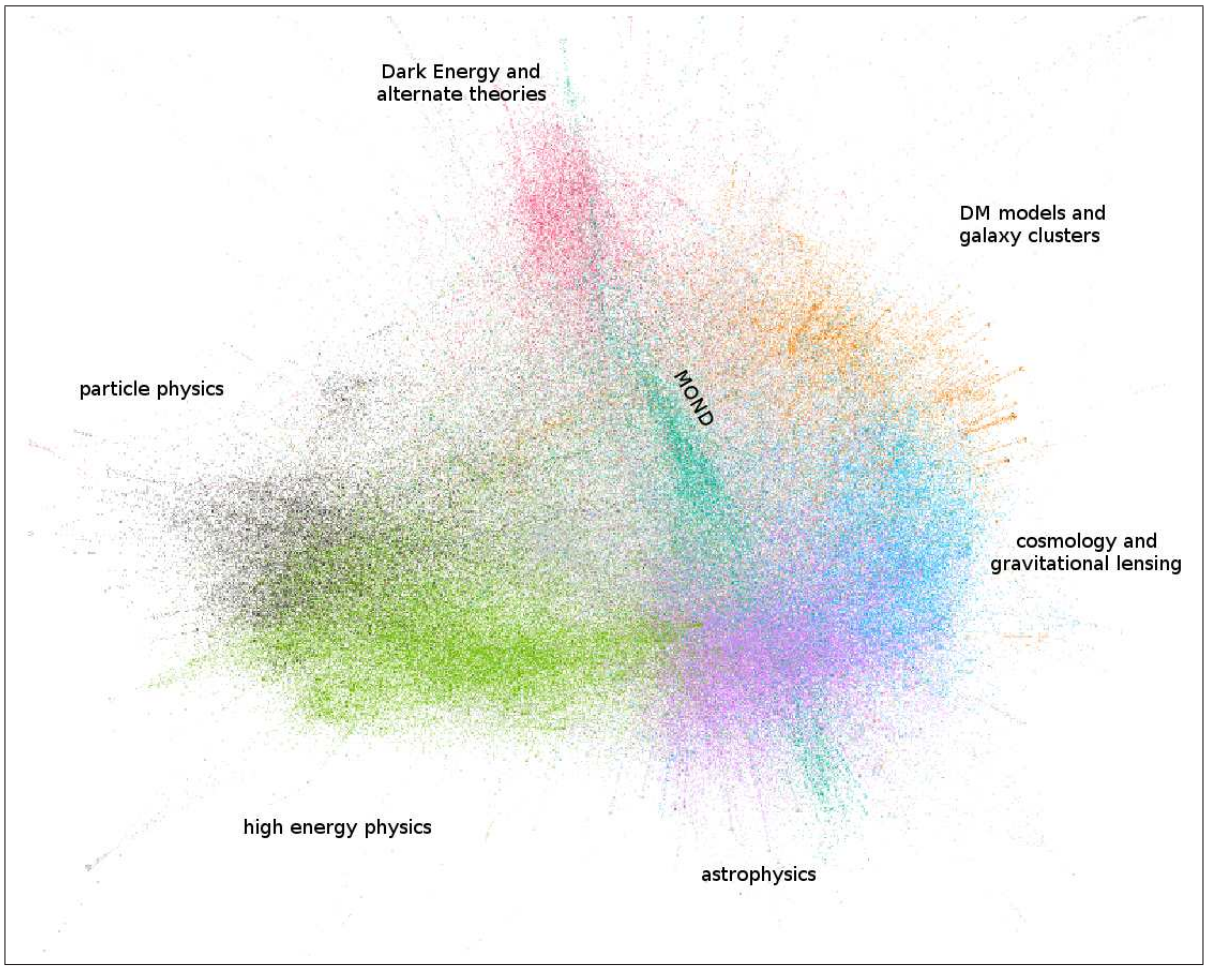

Figure 5.9: Communities of the Dark Matter network (limited to shell 1) visualised using the Force Atlas algorithm, nodes coloured to show overlap of communities, links not shown

others. It bisects the largest community, astrophysics, and is perpendicular to the axis from high-energy physics to cosmology. The overall impression is that a subject matter citation network is condensed by the amount of interlinking between communities, which blurs the distinction between communities.

Evaluation of community-finding algorithms has been assisted by networks where there is an indicator for the *ground-truth* of the network [Yang and Leskovec, 2012]. It requires labels being defined for each node of the network, but there have been

difficulties in equating it with the community structure [Hric et al., 2014].

In the Dark Matter network, there is no obvious label to apply. The arXiv categories are not indicative with many papers concerning particle physics being lodged in the `astro` category, instead of one of the `hep` categories as might be expected. This may be the case in a subject with a large degree of cross-over between fields, but the result is that it is not possible to externally identify communities on this network using categories chosen by individual authors.

To describe the communities of the citation network, labels consisting of the top seven keywords of the group members were applied to the communities and listed in Table F.3. As mentioned in Section 5.3, this method of labelling is not particularly enlightening. Missing keywords are one of the limitations of using this method. Of 148 092 papers with metadata, three-quarters of those (108 660) have keywords. The number of papers with keywords that are definitive in their categorisation of a paper drops to 50% (74 236) of the total. It is reassuring that at least 57 998 papers (40%) include a keyword (from one of the following: cosmology: theory, cosmology: observations, particle-theory and field-theory models of the early universe, high energy physics - theory, high energy physics - phenomenology, high energy physics - experiment or mond) that matches a category of interest. If keywords were the only source of metadata available, then maximising the utility of as many keywords as possible would be attempted. One method would be to use a supervised learning technique, such as a Naïve Bayesian classifier on the 57 998 papers with known labels to extend the number of keywords that could categorise a paper [Nisbet et al., 2009]. Fortunately, the ADS provides other sources of metadata.

In an effort to clarify the significance of the keyword labels, the titles for all papers in a citation network community were examined as a group. Three to five communities were initially selected from each of four ranges of community size. As the number of titles increased, themes were tallied on a sheet to give an indication on the makeup of the communities. Group names given here refer to the Group IDs found in Table F.3.

In the small communities with less than 9 titles, it is quite easy to consider all title at a glance and extract the common themes. Even with two titles, there can be a

range of similarities. Both papers in **Group 587** deal with models in particle physics, but focus on different types of particles, one on the Higgs boson and the other on muons. **Group 808** is concerned with Dark Matter and the Anthropic Principle, the same topic by different authors. **Group 683** consists of a pair of papers by the same authors on the question *Do Active Galactic Nuclei convert Dark Matter into visible particles?* and a follow-up paper with its resolution. Similarly, **Group 822** consists of 2 papers that both have *black hole* and *dark matter wake* (a curious term) in the title, by the same authors in the same year. Slightly larger groups were examined and found to also focus on a single topic. The three papers in **Group 263** all refer to Dark Matter and charged particles, although two of them use synonyms for particles. They have one author in common and were published over an 18-month period. Mentioned in Section 5.3, the question over the significance of the keyword white dwarfs in **Group 4585** now becomes clear. The titles place the keyword in the context of Dark Matter candidates (see Appendix E) with some titles answering questions raised by other titles. **Group 5960** has 5 papers with only the general concept of structure in common. They are published over a 10 year period (1995–2005) by different authors, which explains why they have a lower text similarity amongst themselves than, as shown in Figure 5.3, groups **587** and **683** which both are smaller, tightly focused on particles and are published within 15 months of each other (2008–2009).

Medium communities with 10–15 titles take longer to assess because there are more possible themes to compare with each title.

Group 5955 contains *Brane theory* in most of the 11 titles, yet brane is only the 5th most common keyword in the group. Brane cosmology is a model in theoretical astrophysics involving higher dimensions.

Group 5940 is a more mixed group comprised of traditional cosmology topics with half of the papers mentioning redshift.

Group 5967 is very well described by the particle physics term *axions*, regardless of the other keywords in the label in Table F.3.

Group 811 has more papers with the term *weak lensing* or *weak shear* in the title than it has papers with keywords.

The complexity of assessing large communities consisting of 50–100 titles is immediately apparent when the titles do not all fit onto one computer screen. Multiple passes are required as the significance of terms is not immediately realised.

Group 544 is on particle physics and at least half of the papers have *sterile neutrino* in the title followed by many others considering some form of Dark Matter model (hot, warm or cold). It is noteworthy that 88% of these papers were found in the `astro-ph` category, whereas only 12% were submitted in the `hep` category, which would be expected by the topic expressed in their titles. Despite featuring heavily in the titles, neutrino is only the seventh most common keyword in this group.

Group 700 is concerned with galaxies and data-driven modelling of galaxies based on observations. Mixed in with 53 mainstream papers are three papers on MOND. In this case, the ranked keyword label is an accurate description of group as well as the data in Table F.3 which estimated the membership to consist of 5% MOND papers.

Group 914 concentrates on microlensing surveys and (mostly baryonic) Dark Matter candidates.

Group 5973 consists of topics all in high-energy physics: cosmic rays, particle annihilation and detecting signals of Dark Matter. This is another example of the difficulty in using arXiv categories to identify communities. The papers have been categorised as half `astro-ph` and half `hep`. The keywords are fairly descriptive of the group. Without domain knowledge, it would be difficult to spot that the titles in this group were related.

Very large communities consisting of more than 100 members start to generate more topics than will fit into human working memory. A proper assessment could

require several passes through the titles just to establish which topics are suitable criteria for categorisation of the group during the final pass. Any papers that do not fit an immediately recognisable theme begin to get lost in the cognitive noise.

Of the 109 members of **Group 691**, over 90% have titles that are MOND-related. This is more than discernible from the keywords or arXiv categories. The remaining papers propose tests of gravity, arguably related to a topic which proposes a new understanding of gravity.

Group 739 has 255 members and represents mostly theoretical work (cosmology, Dark Matter with Dark Energy), with about 5% on MOND and some work on particles. The keywords quite representative and the arXiv categories are mixed (9% hep, 1% MOND, 18% other). The large size of the community means that there are small sub-clusters away from the main topic. Dark Energy was easy to spot, but was difficult to keep track of the other topics and it required a concerted effort to deal with this many members in one group.

Group 92 has 476 members dominated by High Energy Physics (`hep`), focusing on Dark Matter searches (both astronomical and particle), detectors and projects, and how the results affect models (both cosmological and supersymmetry) Some papers contain suggestions for Dark Matter candidates (mostly non-baryonic) and host some very imaginative Dark Matter scenarios, such as volcanogenesis. The keywords in this community are accurate, but not very descriptive.

Group 213 has 642 members accurately described by keywords and arXiv categories. These papers represent the traditional astrophysics approach to Dark Matter, and could be considered the main body of work on the subject. An even mix of observational evidence, simulation, theory and models, they mostly deal with galaxies and large-scale structure and distribution of matter. Many papers concern galactic haloes (an observed structure) or gravitational lensing (a detection technique).

While some journals, such as *Monthly Notices of the Royal Astronomical Society*, (MNRAS) *Astronomy & Astrophysics* (A&A) and *The Astrophysical Journal* (ApJ) require that authors select keywords from a controlled vocabulary (a list of pre-defined terms), no such restriction is placed on authors submitting papers to arXiv.

Coverage is another advantage of using titles over relying on keywords as almost all papers have a title recorded in the ADS (148 031 out of 148 092 papers).

Group 811 provides an excellent example for the use of the community finding algorithm where the titles indicate a more specific topic and with greater coverage than the keywords. The most common keyword is *gravitational lensing*, of which *weak lensing* is only one of several possible strands of inquiry. Here, keywords summarise some of what might be encountered in this group of papers, but the titles confirm that a very strong connection in the topic is shared by papers linked structurally in the citation network. This finding of ground-truth matching network structure was not evident in Hric et al. [2014] and represents an interesting new line of investigation.

5.7 Summary

There was some overlap found in the methods drawn from both complex networks and text mining, despite the differing natures of the two aspects of academic papers. It was found that citation communities have a higher textual similarity and can cluster papers into similar groups as partitioned through textual means. The titles of papers were found to be an effective method of identifying citation network communities. A natural size of four text clusters was indicated by the network community finding algorithm. The only effect of increasing the number of words in the stopword list was to lower the average cosine similarity values, not to alter the text clustering. Finally, it was found that the document graph (formed by linking together papers that exceeded a threshold similarity value) does not begin to resemble a complex network in degree distribution until most of the links are eliminated at a threshold of at least 0.6. These first steps in combining techniques have offered insights previously only hinted at by assumption and intuition.

6 General Discussion

In addition to finding that the textual similarity within a community is greater than it is outside it, I found the following results. The Dark Matter citation network has a degree distribution, $P(k)$, where fitting $k^{-\gamma}$ to large k yielded an exponent, $\gamma = 2.3$. An examination of the degree-degree distribution, $P(k, k')$, via degree pair correlations exposed two distinct behaviours for citations. While learning the intricacies of text mining, an upper bound was placed on the error in creating a Vector Space Model due to improper normalisation, the derivation of which is in Appendix D. After processing the documents and computing their cosine similarity, it was found that over a decade the similarity decreases by 10% of its value. Text clustering with k-means was found to have four natural clusters.

Creating citation networks of this type can be extremely time-intensive, although once available they can be used to evaluate newly proposed models of citation networks. The incremental layering of citation shells upon the network has demonstrated a change in the degree distribution. Communities found using the Louvain algorithm have a higher textual similarity between papers within the community than to those external to it. Labelling these communities using the titles of the papers in the communities was found to provide more clarity than the keywords supplied with the papers. In treating document similarities as a graph, a complex network was found in the document graph for high threshold values for the similarities.

Given the document graph is a complex network, it can by analysed with network techniques. Even though the document graph and the citation network are different, as shown by the k-core diagrams of both networks, citation and text networks can be combined in what is called a "bipartite graph", an area that has recently been receiving more attention. The change in degree distribution is a transition from a stationary network to a growing network, one that adds links faster than linear in the addition of nodes, indicating a phase change in the attachment model [Dorogovtsev and Mendes, 2003, p. 13].

The practical considerations arising during the text mining process are that comparing equal numbers of subsets in two partitions is not required for evaluating partition similarity with the Rand Index, that labelling clusters requires care in order to be meaningful and that using citation communities is a novel approach to uncover the natural number of text clusters.

6.1 Networks

6.1.1 Citations and Graphs

The ideal description of citation networks is that they are directed acyclic graphs that grow over time by adding links which are then static. There exist rare conditions in which this ideal is violated in small sections of the network. Prior to online publishing, once a paper was printed in a journal its position in the network was effectively crystallised. Even then, papers could be withdrawn, retracted because of errors in the method, hopefully before it garnered many citations. One of the papers selected from arXiv was removed by a third party on discovery of plagiarism committed by the author in another paper. It and any links to it were removed from the network before any analysis was done.

An extreme example of the effect of pre-prints in deviating from the ideal is *Contacts and Influence* by Pool and Kochen [1978], which was published 20 years after it was first written. Citations to pre-prints can still propagate due to "citation copying" from one list of references to the next, unaware of the paper's change of status [Simkin and Roychowdhury, 2005]. Incidences that can arise from the lag between being made available and being published range from two identities for the same paper, requiring care in resolution, to older papers citing younger papers, seemingly contrary to causality. References can be added during the publishing phase which may then lead to a cycle in an otherwise acyclic network.

At this fine resolution, the actual generation of the citation network is not atomic

or single stepped, but selecting only published papers misses pre-prints and other material that are never published, yet widely available online. One possible refinement developed for graphs, but not yet used in citation networks, is the weighted graph, which ascribes a quantity to a link. Drawing on the work of Teufel et al. [1999, 2006] in argumentation would be a good starting point for extension as would examining the impact factor of the journals used, but the sheer amount of data is a barrier to this approach.

6.1.2 Technical Challenges

This citation network has been produced with considerable effort, but it has been worthwhile to create a substantial example independent of the ISI, a rare commodity. The activity of creating the network has informed the conceptualisation of how the citation formation takes place and pitfalls hidden by the mass of data gathered. Two bugs in `Astro::ADS` were found and rectified, benefiting future users of this publicly available search interface. It is unfortunate that the network itself has not been allowed to be publicly released as it requires anyone other than the ADS to duplicate the effort to verify or extend this work. However, they would have the results available for assessment on whether the outcome merits the required effort to produce it.

A few remarks should be made regarding the technical implementation of the data collection and processing. The retrieval of the data over the Internet was planned to take place over a couple of years. A MySQL database was used to store the data as it was collected. This technique was robust over the long term in the face of network and power outages and possessed the standard tools for manipulation, backup and retrieval, but it was slow to extract the data for processing after 10^5 bibcodes had been collected. A possible solution to the slow extraction is to store the data in a graph database which seems naturally suited to the problem domain [Robinson et al., 2013]. Development on graph databases is relatively recent in comparison to relational databases. Their query languages search for patterns based on the relationships between data, which promotes network exploration and allows advanced questions to be posed concisely. The

effectiveness of this approach, although having had success in some business domains, would still need to be evaluated for use with complex networks.

Algorithms for finding the shortest path between all nodes, such as Floyd's and Dijkstra's, use matrices to hold their results meaning that the memory requirements scale as the square of the number of nodes [Neapolitan, 2003]. As the Dark Matter network approaches 10^6 nodes, these algorithms require up to 10^{12} memory locations. A trillion memory locations is more than is usually available in standard computers and as a result, the algorithms run out of memory before they can complete. To avoid this obstacle, either a very large memory-mapped file is required or a breadth-first search–though the searches are inefficient in time.

Labels constructed using term-frequency ranking of the keywords in the text cluster give an indication of the focus of the cluster, but lack clarity. It is simple to understand what topics are covered by the cluster, but not to understand the central direction of research activity in the cluster. As mentioned in Section 4.9, producing good labels is intensive work. It is alike to the subject of Topic Modelling in its desire to succinctly describe a group of documents [Crain et al., 2012].

Topic Modelling is a technique of dimension reduction which statistically analyses document collections to discover the latent topics contained in the collection. It depends on building a predictive model of terms that stand out in the text, usually using LDA, a bag-of-words representation [Blei et al., 2003]. It requires care and effort to develop such a model and methods for doing so are being improved as limitations are being discovered and addressed [Blei and Lafferty, 2007]. Current work by Tang et al. [2016] uses a random walk on the topic graph to achieve data summarisation. Such summarisation applied to a document cluster should satisfy the requirement for labelling the clusters, provided the researcher has the time and expertise to build a Topic Model for the corpus. Huang et al. [2016] has tried to find topics without the requirement of prior knowledge of the anchor words in the subject by finding the most selective topic words from the corpus, but some expert knowledge is still needed to validate the topics found.

6.1.3 Selection Criteria

The selection process for the papers included in a citation network is normally based on all papers published in a selected journal or a broad category within a journal or database. Networks that include all papers in a journal encompass all topics catered to by that journal. Likewise, many topics are included within a category. Selecting many topics can have an averaging influence on the nature of the network. Even the Dark Matter network, based on a single topic, draws in communities with very different publishing cultures, from lone theoreticians to large-scale high-energy particle experiments with a multitude of authors.

It is suspected that the differences in degree distribution is more likely due to the selection process than the choice of database, but either reason, if verified, would be illuminating. A discrepancy between the data sources would be of extreme interest to the database owners.

6.1.4 Network Clustering

Clustering is an integral part of a complex network. The Dark Matter network has a clustering coefficient of 0.216, much greater than the value expected from a random graphs of this size, where $\bar{C}_r = \frac{2L}{N^2} = 2.51 \times 10^{-5}$. There is a strong relationship between clustering and node degree. As shown in Figure 3.6, this relation holds at all sizes of the network. The drop in clustering as degree increases is perhaps due to the nature of attracting citations from many different threads, which dilute the number of clusters realised in the network due to the increased number of nodes available, although there is a natural connection between the average degree of nearest neighbours and the formation of cliques. Another argument is that papers with few citations, when cited, are cited together as the core of a sub-topic, a list of results that naturally group together or self-citations for historical context.

6.1.5 Pair Correlation

The degree-degree distribution of Equation 3.4 was investigated via pair correlations. Pair correlations are a feature of real networks and the results shown in Figure 3.10 display a dual nature. At low degree, the plot shows a clear correlation between the in-degree of a node and the in-degree of their nearest neighbours as is also seen in the topology of the Internet. Unlike the Internet, the correlation vanishes for $k_i > 30$. The disappearance of the correlation at $k = 30$ is called the cutoff degree (k_c) for the pair correlations and, according to Boguñá et al. [2003], is due to the finiteness of the network. Could this also be a result of the split between normal papers and very highly cited papers? If so, this value is a good candidate for a parameter to be included in models of citation networks that take account of "super-joiners" as the transition in Figure 3.10(b) makes a clear distinction between the two ranges.

Pair correlation is another facet of clustering in the network, describing the linking behaviour in local neighbourhoods. The monotonically-decreasing section of the degree-degree distribution below $k = 30$ in Figure 3.10(b) is typical of disassortative correlations in growing networks and indicates a hierarchical organisation of the network [Serrano et al., 2007]. Disassortative correlations indicate that highly-connected nodes are preferentially connected to sparsely-connected nodes. This result is corroborated by the k-core visualisation in Section 5.2. It would not be unexpected if citation networks were to be considered in the same category of network as technological or biological networks as opposed to social networks which are assortative by nature (highly connected nodes are preferentially connected to other highly connected nodes). While citation networks are fairly well understood in terms of their clustering statistics, it has not been explicitly stated that they are disassortative in their degree-degree correlation [Velden et al., 2017, Šubelj et al., 2016].

A contrary result is reported by Xie et al. [2016] who examined the DBLP citation network of Computer Science papers from 1936–2013 and found that the average nearest-neighbour degree to be an assortative correlation. There are clearly substantial differences between the DBLP and Dark Matter citation networks such as the field,

range of subjects and time span covered. They calculate a clustering coefficient of 0.070 for the DBLP citation network for papers up to 2013-09-29 whereas the Dark Matter clustering coefficient is 0.216. They also claim to see the assortativity in citation networks in the arXiv categories `hep-th` and `hep-ph` (clustering coefficients 0.165 and 0.148 respectively). Assuming that both results are valid, that the DBLP citation network—with clustering coefficient 0.070—is assortative and the Dark Matter citation network—with clustering coefficient 0.216—is disassortative, it might be reasonable to expect a turn-around point to lie between the two clustering coefficients for citation networks. If true, it would be disastrous for their 2016 model which produced an assortative citation network with a clustering coefficient of 0.390.

Models that neglect degree pair correlations fail to recreate real networks [Serrano et al., 2007]. If a case is to be made that there are two classes of citation networks, assortative and disassortative, then the defining characteristics that change the dynamic from hierarchical to social need to be identified. Possible factors for examination are the degree distribution, growth rate, modularity, average degree or multi-disciplinarity of the network. Ravasz and Barabási [2003] argue that a scaling law in the clustering distribution of

$$C(k) \sim k^{-1} \tag{6.1}$$

quantifies the coexistence of a hierarchy of nodes and as such should be observed as the clustering coefficient changes. The prominence and location of the phase transition will be useful in evaluating and improving on generative citation models such as Golosovsky and Solomon [2013] and the follow up work by Golosovsky and Solomon [2017] which, while not referring to phase transitions, continues to incorporate two categories of paper with different ageing patterns. These models are applied to the question of how researchers choose their citations. The cause of the discrepancy in assortativity between the two citation networks merits further investigation to improve models and understanding of the dynamics of the act of citation.

6.1.6 Growth of Networks

The growth of the citation network is not uniform across the degree distribution. Comparisons of citation networks of different sizes have always come from different sources, not from the same network as their selection criteria has been journal boundaries, for the most part, which precludes outward growth. Starting from the core Dark Matter network of the Dark Matter corpus and all of their references and citations and comparing it to the whole network consisting of all of the references and citations to the core network, it is shown in Figure 3.8 that, not only does the power law exponent increase to a value closer to Redner's citation network statistics, but also develops a hump also seen in other citation networks. The position of this hump near $k = 45$ adds a scale length to what was previously a scale-free network. Golosovsky and Solomon [2013] labels this a dynamical phase transition in terms of microscopic growth models, separating papers that have a citation lifetime of 6–10 years from those that have a much extended lifetime. The increase in N by a factor of 10 has pushed the distribution towards higher degree nodes as demonstrated in Figure 3.9 with the shear of the curve to the right. The implications of the embryonic network are not clear.

6.1.7 Significance

How does the lower value for the power law exponent fit into the context of research on citation networks as well as its significance to Dark Matter communities? The exponent γ needs to be used in conjunction with the clustering coefficient and the degree-degree distribution (discussed in Section 6.1.5) to produce improved citation network models. These statistics are the building blocks used to construct generative models which are then tested against real citation networks. Validation of models requires data, meaning that more citation networks need to be collected and their statistics measured in order to improve the models. It is an iterative process.

The significance of the Dark Matter citation network is that it is a rare specimen of a citation network not defined by the boundaries of a journal or subject, but built

around a topic. It is potentially useful in examining the citation dynamics of a more homogeneous research community by controlling for localised variations in citation practice. With a good network model for a single topic, the networks based on journals can be modelled as a superposition of parameterised single-topic models according to identified topics.

The power-law exponent, γ, was found to be 2.3. That this is lower than values found by Redner [1998] raises several questions. A lower value signifies a larger ratio of papers with more citations to papers with fewer citations. Is this an artefact of the construction method? The overlap of communities studying Dark Matter as seen in Figure 5.9 is one influence on the low value for γ. The distinctness of topics in a journal-defined network implies fewer connections across unrelated papers although the clustering coefficient is comparable [Šubelj et al., 2016]. Is this a deficit of heterogeneity in the citation conventions of distinct research communities? Only by constructing more single topic citation networks can this question be answered. In gathering these data sets, the common features among them hold the promise of a citation network model for single topics which could be used as building blocks for more complex citation networks.

6.2 Text Mining

While a soft clustering technique such as Expectation-Maximisation is a natural choice for clustering documents that have overlapping content across communities and, at times, belong to two or more different categories in arXiv, an algorithm for comparing soft clusters was only published seven years ago and no implementation was available at the time of the investigation [Campello, 2007]. Had it been available, it would have been interesting to examine the relation between textual overlap as measured by the soft clustering and the numbers of citations between communities.

The Dark Matter corpus is a narrow range of documents. To understand its character the distribution of the similarities between the documents were plotted and

Figure 4.1 shows them to be skewed to lower values. It demonstrates that abstracts are less similar than the text body. They are much shorter than the main body of text and fewer words have fewer opportunities to overlap. The distribution of keywords is flatter. They are specifically chosen to represent their document and while not exactly drawn from a controlled vocabulary, the limited number of keywords in use degrade the statistics of the distribution. PACS numbers are used for standardising keywords for common topics but they are not universally used and are not immediately apparent to the reader, even though they solve the normalisation problem (see Section 4.3.3) for keywords. The effect of increasing the stopword list is to reduce the overall similarity, a small reduction of noise in the clustering algorithm.

While based on a wide spread of measurements, the fit in Figure 4.2 indicates a decrease in similarity of 0.01/decade. The small change over time is reasonably the result of movement in the topic. The linear fit is only the general trend of the data and not yet an accurate model of the affect of age on document similarity. The size of the decay in similarity is small compared with the spread of similarity values. The number of measurements is the reason that any statement can be made about the ensemble with any confidence. To substantiate the claim for a real effect, the data was cast as a box and whisker plot (Figure 4.3), showing the median and interquartile range plotted with the line of best fit to the data. It is clearly seen that the medians lie close to the line, falling evenly to either side, demonstrating a strong linear tendency over a 15 year period.

Given the error estimate from the derivation of incomplete normalisation, it is possible to arrive at an order of magnitude for the change in vocabulary over time in the Dark Matter corpus. With a change in the average similarity (0.514) of 0.006/year, the relative change in similarity is 0.0117. If the relative change due to a single term change in a small document is estimated at 0.001, than the trend of the change in the Dark Matter corpus is the order of 10 terms/year. This value for the vocabulary change over time could be used in comparison with other diachronic studies (discussed in the next section) to estimate the speed of scientific innovation against the background of the general evolution of vocabulary.

6.2.1 Text Clustering

It might have been unfortunate that the k-means clustering did not provide a clear indication of a natural number of clusters, but that lack may have prevented a strong claim for a number of fixed communities offered weakly by other algorithms that *do* try to determine k. A contradiction of that point is raised by the partition similarity plot in Figure 5.5 which reveals a strong peak at $k = 4$ when compared to the communities in the network found by the Louvain algorithm. It confirms that a broad breakdown of Dark Matter into the four groups of observational astronomy, theoretical astrophysics, high-energy physics and MOND are valid, as well as how there are finer distinctions that can also be made.

Of the possible cluster sizes inferred from the $\widehat{RSS}_{min}$ graph in Figures 4.4 and 4.5, $k = 4$ was only marginally more distinct than the other values. The gradual rise seen in all Rand plots is assumed to be because it is easier to be more similar when more clusters are available.

Error estimates in similarity measurements (as derived in Appendix D) are not often considered; heuristic approaches being the preferred method of validating results. The calculation of incomplete normalisation of the VSM representation of a document holds more utility than just reassuring the user that small errors can be disregarded. It is an estimate on the granularity of the similarity values, setting the threshold of significance below which two similarity values can be considered equal. Using a derived estimate based in theory provides direct accountability, a value open to inspection and discussion. It is portable between data sets and simple to apply consistently.

A side-effect of determining the scale of similarity-space is considering the appropriateness of hard or soft clustering. Section 4.7.3 shows an example of using this scale to count the number of documents that lie close to the boundaries of the k-means clustering of the Dark Matter corpus. Having found that only 7 out of 2659 papers lay close to a boundary, the assumption that papers can reasonably be assigned to a main cluster or category is claimed to be valid in the context of hard clustering. This claim does not rule out secondary characteristics or multi-cluster membership for the papers,

only that a primary cluster could be found for almost all of them. If more papers, say 10% of the corpus, were near the cluster boundaries, a case for hard clustering becomes more difficult to make and the multi-membership model of soft clustering is called for. A deterministic establishment of a scale for clustering will inform algorithms that make use of granularity in the clustering process, such as rough k-means [Peters and Weber, 2016, Huang et al., 2014].

A possible avenue for investigation of the scale of similarity measurements is in the analysis of the change in vocabulary usage over time, also known as diachronic linguistics [Bybee, 2010]. Rather than tracking the evolution of individual words, the error estimate generalises the magnitude of change in language in a large ensemble of documents. This captures the motion of the language and the progression of a topic. To better isolate topic progression, it is necessary to subtract the change in language usage over time. Hamilton et al. [2016] has quantified the polysemy of words over time to study the semantic change of innovation and conformity in language. The knowledge of the natural change in language over time combined with the scale of similarity change effected by individual terms should make it possible to quantify the speed of progression in the topic. The progression speed characterises the research activity, highlighting interest and focus, a recurring theme for recommender systems [Kim and Chen, 2015].

Exploration into the deeper significance of the incomplete normalisation derivation is merited both mathematically and experimentally. For a single term, the error estimate varies as the square of the term frequency of the unnormalised term ($err \propto g^2$). How many more terms can vary before higher order effects must be considered? Consider the effect on the derivation of the general case of the similarity between two documents rather than a two slightly different copies of the same document and how the error varies with similarity. At what point does a mathematical derivation become intractable? Experimental modelling using Monte Carlo simulations of virtual document collections constructed according to Zipf's law and the statistical laws of semantic change [Hamilton et al., 2016] could reveal the effects of the number of terms that vary over a document collection.

While suggestions have been made regarding possible lines of enquiry stemming from these results, the feasibility of extending these suggestions to other application has not been determined. Any claims to utility should first determine that they have not duplicated previous results and that they demonstrably improve understanding of the problem space to which they they are applied.

6.3 Communities

Given the selection criteria, using arXiv categories is not a viable method to specify distinct communities, as 85% of papers belong to the `astro-ph` category. Instead, these communities have been found using only the citation data and yet there is a textual overlap amongst the documents that is quantifiable. The results presented in Section 5.3 are evidence that there is a small but statistically significant increase in the textual similarity between papers within the citation community compared with papers outside the community, answering the first research question. The difference in the distributions is highlighted in Figure 5.4, showing how much the average textual similarity (as calculated by Equation 4.2) increases for documents within citation communities. The significance of the difference between the two distributions is that it confirms that on average, citation communities share a higher textual similarity than they do with other papers in the same subject. It is an assurance that continuing research in this direction (such as Mei et al. [2008] in Topic Modelling) merits the time and effort required. The result is not a given because of the narrow focus of the subject. Unlike a corpus of news articles encompassing many different areas, some similarity in content and style is expected. Also, the demands of academic writing dictate that articles cannot be *too* similar. Taken together, the range of similarities contracts, effectively pushing papers outside the community closer in similarity to papers inside the community. The heat map visualisation of the data quickly identifies the cases of interest which may be obscured by the large numbers of comparisons to be made. It is immediately apparent that one group is different from all other groups. No group or

pair of groups has a very high similarity averaged over all documents, although individual pairs can exhibit such high similarity. This spectrum of overlap is an indication of heterogeneity, not a monoculture. In Figure 5.3, the mean of the average similarity between groups is 0.25 ± 0.11 and a similarity of 0.6 is seen only between five papers in two citation groups, a rare occurrence. The highest threshold in Table 5.2 corresponds to the number of links comparable to a citation network, whereas the lower thresholds have more links than other comparable sizes of networks. [Dorogovtsev and Mendes, 2003, p80]

Understanding the groups is not as simple. Labels can be applied to groups (see Table F.3), but many of the keywords are the same and only differ in how prominent they are in the label. It is difficult to extract deep meaning using an automated process. Part of the difficulty is that keywords were found mostly to signal secondary topics of the paper and, in some cases, fail to cover the main topic indicated by the paper's title. Some authors may be using title in conjunction with the keywords to make their paper searchable by the interested reader. As a result, the keywords can be an incomplete description of the topic of the paper.

Identifying communities through the titles of their papers (described in Section 5.6) is a more powerful method, though it comes with its own challenges. Some domain knowledge is required to resolve synonyms or to identify proper nouns, such as project names. Examples of tacit knowledge in the domain include notational shorthand for *redshift* is denoted with a z, where a *high* redshift is $z \gtrsim 2$ and that *brane theory* is a high-dimensional model of space-time which involves 10–12 dimensions. Given a small degree of domain knowledge, no small community in the Dark Matter citation network was unclear in its topic nor did it include off-topic papers. The labels derived from titles that were attached to the seven largest communities found on the citation network were used to identify the communities in Figure 5.9. The figure shows the ties and overlap between the different communities and how they can be interpreted on different scales.

This method requires more cognitive effort as the size of community increases. Identifying the different topics within a community by inspection is a related task to

the knowledge elicitation technique of card sorts. Rugg and McGeorge [2005] find that 20 or 30 is the maximum number of entities for a conveniently manageable card sort. This effect was observed in increases of task difficulty as the size of the inspected community progressed to large groups (50 members) and again with extra-large groups (100 members). As an aside, the range in sizes of citation communities found by the Louvain algorithm provide a stable dataset with which to explore sample size viability for card sorts.

Recent results by Hric et al. [2014] show that 11 different community finding algorithms, including Louvain, do not necessarily recover the ground-truth extracted independently from many large networks. Despite this, there is a higher average textual similarity between papers inside a citation network community than between papers in different communities. The average similarity inside communities is 0.34 ± 0.12 while the average similarity between communities is 0.25 ± 0.10. Student's t-test for significantly different averages indicates that there is a 99% probability that the difference between the two distributions is significant (the null hypothesis is rejected, $t(452479) = 458.87, p = 0.01$).

Fortunato and Hric [2016] explain the reasons why ground truth from metadata may not map well onto structural communities. The application of Text Mining on the Dark Matter network shows one possible way of establishing a better ground truth by using much more information about the nodes on the network. Knowing now that textual similarity is associated with citation communities, more advanced Text Mining techniques can be sensibly pursued. Topic Modelling involves identifying topics and calculating the distribution of terms for those topics across the corpus. The distillation and measurement of a document via Topic Modelling has the potential to capture a more complete representation of the document than the currently available metadata about it. Evaluating whether a topic model provides a better ground-truth than metadata is a risky proposition without the assurance of textual similarity with the communities.

A possible use of these values of similarity within a community is to highlight papers which are textually very similar (explicitly above the average similarity within

a community) that are *not* structurally related. These papers are of interest because they are working in isolation (through ignorance or choice). If the authors are unaware of each other, could they progress further or faster working together?

Use of the Louvain algorithm has been very instructive in analysing the field of Dark Matter. By finding communities of different sizes from small to very large, the investigation of those communities has revealed the major themes, some of the topics of interest and some specific questions. This "top-down" view of the field is both informative and provocative, providing enough background for a researcher to start speculating on how these communities relate.

A desired outcome of identifying the network communities is a simple and meaningful label to describe in some way the majority of the members of the community. Without a clear and natural separation between communities, this task can be non-trivial. After all, the basis for the construction of any connected network is that each member has a relationship with at least one other member of the network. Commonality is inherent.

Because many of the papers in the Dark Matter network have keywords associated with them by their authors and made available by the ADS, keywords are a natural approach to labelling communities. By ranking all keywords in a community from the most frequently found to the least, a descriptive label can be applied to the community. This method is simple, consistent and requires no expertise, but may lead to labels which are opaque and lack the clarity of meaning. Keywords are used by authors to quickly advertise the main themes of their paper. The paper itself explains how the keywords are related; the connection between them is not immediately apparent. As a result, the general idea of the community can be conveyed by keywords but it lacks coherent meaning. This is compounded by the all or nothing nature of keywords for a paper. No indication is given of how relevant each keyword is to each paper, except perhaps by the order in which they are listed. Minor topics may be over-represented in each paper, but a large number of papers in a community should minimise the over-large influence of minor topics.

Another technique for assigning labels is to examine all the titles for the papers

in a community. This requires some expertise to identify the significance of terms, but as a whole, the title communicates a complete idea and the main topic of the paper. Keywords are also used to flag ideas that are not included in the title. It was found to be easier to summarise a list of titles and find the common themes than to construct a label full of meaning from the ranked list of keywords. The disadvantages are that it is both labour- and knowledge-intensive to undertake this task and as the number of titles in a community exceed a single computer screen height, the cognitive burden of summarising the titles becomes heavy. Section 5.6 discusses the specific difficulties involved.

During the community labelling process, some insight was gained into the topics within Dark Matter through their titles. By no means a comprehensive list, any attempt at Topic Modelling should consider the following terms in conjunction with the keyword labels for communities in Table F.3:

keywords listed in Table F.3 clusters, cosmology, gravitational lensing, haloes, modeling, particles, structures, SUSY (supersymmetry), wavelengths (X-rays, gamma rays, radio waves, etc.)

keywords listed once or twice axions, branes, dark energy, detection, distribution, MOND, planets

not appearing in Table F.3 CDM (cold dark matter), WDM (warm dark matter), gravitational waves, terms related to dark energy (cosmological constant, Lambda (Λ) and quintessence), Dark Matter searches and projects

Identifying terms belonging to the categories *searches* or *projects* will require domain knowledge, although some heuristics (names in all uppercase, such as PAMELA) are available. The application of this to the scientific publishing industry is as a method to find new keywords as they emerge in a constantly evolving research environment or to assign missing keywords to papers with minimal human intervention. As a method of keyword discovery, it is much easier to present experts with a collection of titles from a small citation network community and ask the question, “What do these papers have

in common?", than it is to ask experts to generate new keywords. It is a cognitively easier to spot differences and similarities in something presented to an audience than for them to imagine instances without priming [Kahneman et al., 1982].

Finally, the results are mostly consistent with the initial assessment of there being at least three or four large communities within the subject of Dark Matter, consisting of astrophysics, high-energy physics and MOND (mentioned as the motivation in Section 2.1 with a description of the problem in Appendix E). It was not known if the communities on the citation network would recover any of those or if new ones would be found. At the largest scale of the Louvain community finding, four communities were found split roughly into observational astronomy, theoretical modelling, high-energy physics and MOND. No unexpectedly new communities were found. Louvain's first pass finds a distribution of community sizes that reflect tighter focus on topics within the four overarching communities. Four for the number of communities is also supported by the k-means clustering in Section 4.7.3. Some of the topics revealed a focus on the various experiments in high-energy physics to confirm or refute the existence of sub-atomic particles predicted from theory. Louvain has exposed areas of interest, topics large and small that could be confirmed by experts by requiring effort to compile, risking the omission of topics due to the number of communities. In this way, it is certainly a useful tool for review authors, chosen for their broad knowledge of the subject, but still liable to missing small topics outside their experience. Louvain collects topics in a community and its sizes provides a scale for judging the effort to allot to reviewing it.

6.3.1 Document Graph

Having already treated document similarity as a graph created by linking documents which have a similarity above a given threshold, it seemed natural to investigate whether the document graph has the characteristics of a complex network. The results of various thresholds, presented in Table 5.2, show that the average clustering coefficient exceeds that expected in a classical random graph of that size as well as

that of the citation network and the average path length, $\langle l \rangle$, is always less than 10 even when the number of available links to traverse is reduced by several orders of magnitude. The third component of a complex network is a fat-tailed degree distribution which is certainly present at all thresholds (Figure 5.7), but only the thresholds at 0.5 and above show any similarity to a power law relation. The exponent fitted to the 0.6 threshold is 1.20 ± 0.06, lower than required for a growing network.

6.4 Summary

The studies described above show that there is more textual similarity between papers in a citation community than with papers outside the community. There were differences in the degree distribution between the Dark Matter network and previous work, which changed as the Dark Matter network grew in size. A dual character in the degree-pair correlations was noted with implications for constructing models for citation networks. Also with the degree-pair correlations, the network was identified as being disassortative, not previously reported in the literature. Using two clustering methods indicated that the natural size is four groups for text clustering with k-means in the Dark Matter corpus. This finding is supported by the identification of four subjects in the Dark Matter citation network, observational cosmology, theoretical astrophysics, high-energy physics and work on alternative non-Dark Matter models, such as MOND. Using titles of the papers in the citation network communities was found to be an effective method of identifying the communities.

7 Conclusion

The approaches of complex networks on citations and text mining on documents are both well established. What is novel is in combining the two approaches to better understand the nature of a scientific community. This thesis shows that there is a stronger textual similarity within citation communities researching the Dark Matter problem than there is between those communities, as explained in Section 5.3. The relationship is that network connectivity implies textual similarity above the baseline similarity between random papers in the subject. Given a highly interconnected citation network, it follows that a high degree of textual similarity between any two papers is not unexpected. It would be easy for any relationship to be buried in the noise of so many connections and similarities, but the positive correlation between citation network structure and textual similarity stands out because of the numbers measurements available. This thesis also shows that in identifying citation communities, the examination of titles of the papers in a community provides more clarity than does the aggregation of the metadata (keywords) assigned to the papers. Both of these answers illustrate the potential of combining techniques from both complex networks and text mining.

A number of unexpected findings were uncovered. The power law exponent of the degree distribution, $\gamma = 2.3$, is lower than expected from other studies. It was suggested that the selection criteria for a citation network may be responsible. The degree pair correlation shows that the network is disassortative placing them in the same category as technological and biological networks and the behaviour of the correlation changes as the degree exceeds 30, a threshold value that needs to be considered in building future models of citation networks. The document graph acts like a complex network when only high similarity papers are linked. Document similarity in the Dark Matter corpus decreases by approximately 1% of its value per year on average. It was found that comparing the partitions of unsupervised learning across two facets revealed that four clusters are most likely a natural grouping in the Dark Matter corpus, where

individually no preferred size was indicated. Finally, it was noted that the degree distribution evolves from a static to a growing network as more citations are added.

7.1 Further Work

The importance of this thesis lies in the future progress that can build on these results. Five potential directions are described, but the most intriguing is that of Topic Modelling (Section 7.1.5). It has the potential to solve the issue of ground truth in network communities, but without knowing that citation networks exhibit significant textual similarities, demonstrated here in Section 6.3, the work required to develop a Topic Model is at risk of being wasted.

7.1.1 Testing Models

Having gone to the time and effort of building a large citation network which is unlikely to be re-created, a useful endeavour would be to assess the various models currently available and report how well they describe the Dark Matter network. Most models have been empirically tested, but not against the same dataset, presenting difficulties with comparison. Unfortunately, not all descriptions of models are expressed in the same mathematical terms and care is needed in correctly applying disparate approaches to determine equivalencies. In general, these types of reviews are a boon to progress in any field, but rare because of the effort required to complete them.

7.1.2 Document Analysis

Downloading the documents from arXiv provides access to the source material which produces the documents. In the field of physics, the source of most research papers is written in LaTeX which yields more information about the structure of the document allowing finer control over parsing choices, such as extracting abstracts or equations.

An extension of the themes presented here would be to use LaTeXML to locate citations in their surrounding text to quickly characterise the type of citation in CiTO, the ontology described in Section 4.1.1. After many learning examples, a supervised machine learning technique could try to extract the basic principles in order to automate the process. This procedure is a way of managing the volume of data for large scale argumentation analysis. Subsequent work in this direction should try to augment the work of Teufel et al. [2006] with their annotation scheme for the rhetorical function of citations with regard to supporting or refuting prior work. Current attempts at automating citation identification has been undertaken by Di Iorio et al. [2013] using Semantic Web and NLP techniques.

7.1.3 Terminological Distance

Trying to find how much work needs to be done to successfully enter a new topic as a researcher requires text from outside this topic to measure the conceptual distance between the two topics. Extending from this work, a suitable start point would be to select the papers that are on the edge of the topic identifiable by unique keywords or categories such as `physics.bio-ph` and `physics.geo-ph` in the Dark Matter corpus and extend the citation network in the direction of the references outside the topic until there exist several citation communities found on the other side of that bridge. Using as much text as possible from the communities directly connected by the bridge, remove all common words to mark a baseline and collect terms shared across the bridge. Terms unique to the citing bridge paper will consist of author style, contribution to knowledge and, most interestingly, embedded implicature of the foreign topic. In bringing to the table concepts developed in another topic, new ideas need to be explained to be accepted as argument. Hence, statements which would be implied in ongoing discourse must now be explicitly laid out to the new audience [Grice, 1989]. Thus, the concepts necessary to cross topics are enumerated and described by bridges. It is unlikely that a sufficient number of concepts for a general reading of another unrelated topic will be made available in one such bridge, nor is there a guarantee that the new concepts have

been correctly communicated, for surface features can be misleading to the uninitiated. In today's research climate, the fostering of multi-disciplinarity would be facilitated by a primer of terms compiled from all relevant bridging papers across the fields in question. Future explorations on this theme should attempt to discern the effect of soft unsupervised learning on the comparisons between citation network and document clustering. An implementation of the Fuzzy Rand Index will be required.

7.1.4 Bipartite Graphs

As described by the CiTO ontology, there are three main features that can tie a collection of academic papers together like three sides of a triangle: the authors, the citations and the content [Shotton, 2010]. While the citation network and the co-authorship network have been examined together as a bipartite graph, the same comparisons have not been made with the content for several reasons. Relying on the curation of publications performed and made available by the ADS, citation and author data are relatively uncomplicated to collect and process with clear methods tested against a multitude of networks that routinely deal with up to a million nodes. To collect and process the same number of documents requires an order of magnitude more investment, akin to the activities of commercial search engines. Text comes in many disparate formats which have individual parsing requirements. Because the size of the representation is much bigger, the time to process the document and manipulate the representations impact the ability to match citation or co-author networks. Those two networks have their differences, one being a directed graph and the other undirected, but are both familiar to the complex networks community. The skills required to produce the document graph are distinct, doubling the work required to examine the subject and for uncertain gain. Finally, publisher's restrictions make accessing large numbers of documents problematic at best with no guarantee of reaching all journals of interest or whether the text is available in an electronic format. While Section 29A of the UK Copyright Act 1988 (and its 2014 amendment No. 1372, *Copyright exceptions for research, education, libraries, museums and archives*) permits usage for non-commercial

research, one must first obtain lawful access which the publisher is not compelled to provide.

7.1.5 Topic Modelling and Citation Communities

Attempts to reconcile network communities with ground-truth labels have not met with great success previously in large data sets [Yang and Leskovec, 2012, Hric et al., 2014, Yang and Leskovec, 2014]. In these cases, the labels represented individual's membership to different communities expressed by one facet of their existence. These are the labels that were available and readily accessible. They look at where an individual happens to be in a particular classification, not what attributes community members have in common. There may be some subtle interactions missing because of the bias of availability. Given the relative surety with which the Dark Matter citation network communities were characterised using only the titles of the papers within the communities, perhaps the question needs to be turned around. What labels *should* be inferred from the communities found in a network and how well do those labels reflect their communities?

Topic Modelling has been done with citation networks, but no attempt was made to examine the communities within the citation network [Lim and Buntine, 2014]. A novel approach to the assessment of community detection algorithms on citation networks would be to use a Topic Model (developing it for the corpus, if a suitable one is not available) to label the communities found by the algorithm and evaluate its performance against random clusters of the same size distribution. As discussed in Section 6.1.2, could a more interesting ground-truth be discovered by Topic Modelling, in so far as relates to the structural connections between members? Topic Modelling without citation communities is just looking at what terms are associated with each other and how their evolution over time. Topic Modelling *with* citation communities adds the structural dimension and explores how research proceeds, perhaps expanding on the question of what is required in order for research to progress. Any progress should be illuminating.

If this method is found to be productive, future work should extend it to other complex networks (such as protein-interaction networks) where community detection is used to identify functional groups [Rosen and Louzoun, 2016]. Rather than infer function of unknown members from what is known about other members of their communities, use the techniques developed in anchor-free Topic Modelling [Huang et al., 2016] to expose new connections for study. Even if the method does not provide immediate applications, the investigation should add to the understanding of the different facets of community detection, a question recently raised by a few of the prominent researchers in community detection on complex networks [Schaub et al., 2016].

A Repositories

In the early nineties, researchers started to expand on the idea of pre-print servers creating the beginnings of today's digital research repositories. The power of these repositories is in their ability to search a very large selection of articles and deliver them quickly through a web-browser anywhere in the world. ADS has even produced an API for software such as `Astro::ADS` (see Appendix C) to automate downloads, but these downloads have restrictions in order to maintain the level of service for everyone. In order to download the large numbers of files required for bibliometric research, permission must be sought from the data controllers.

A.1 ADS - the SAO/NASA Astrophysics Data System

The ADS is a Digital Library for Astronomy and Physics, funded by NASA [Accomazzi et al., 1994, Eichhorn et al., 2002]. They collect publication data from all astronomical and related physics journals made available for the use of all researchers via their web interface. As publishers consider citation data as a part of their intellectual property, the ADS has had to agree to restrict large-scale access to the data. An agreement was reached with the ADS Program Manager, Alberto Accomazzi, to download citation data using the `Astro::ADS` interface.

Bibcodes are a 19 character string that encode aspects of the article metadata into a unique identifier for each article indexed by the ADS. They take the form

YYYYJJJJJVVVVMPPPPA

where the fields are broken down as:

- **YYYY** is the year the article was published
- **JJJJJ** is the journal title abbreviation

- **VVVV** is the journal volume
- **M** is a journal type identifier
- **PPPP** is the start page of the article
- **A** is the first letter of the first author's last name

The fields are padded with full stops to prevent empty spaces and the & is a legal character in the journal title abbreviation.

The citation data is not guaranteed to be complete as there are issues with automating the extraction of references as described in Section 3.2.2. There is a form for users who find missing or incorrect references in the data to update the ADS database.

An overview of the project is given by Kurtz et al. [2000]. The architecture is described by Accomazzi et al. [2000], the search engine by Eichhorn et al. [2000] and the data holdings Grant et al. [2000].

A.2 arXiv

arXiv is the most widely used pre-print server, at one point with mirrors in seven countries around the world [arXiv.org, 2014]. It started service in August, 1991 as `xxx.lanl.gov` hosted by the Los Alamos National Laboratory and its initial submissions reflect that origin in High-Energy Physics. It now accepts submissions in physics, mathematics, computer science quantitative biology, quantitative finance and statistics. The articles submitted to it are not peer reviewed, although moderators ensure that papers are relevant to the available categories. Indiscriminant downloads are prohibited because the repository is hosted on a small server which cannot handle large network traffic. To keep the service available to everyone, scripts downloading many articles in a short period of time are banned.

They offer access to PDF and PostScript files, which they generate from the original sources on the server before sending them out. They will also provide a DVI

file or the original source, which depends on the submission format. The original source files are frequently in a more structured format (LaTeX) than the published version.

After discussions with Jake Weiskoff, an arXiv administrator, I was given a tar file of 2671 papers to download which resulted from a search for the term Dark Matter at the arXiv site on 20 November 2008. They provided the source files, mostly LaTeX, with a few PostScript files, PDF, HTML and Microsoft Word documents.

B LaTeXML

LaTeXML is a package that converts LaTeXfiles to XML, assisted with bindings to customise the rendering of different style files. Bruce Miller created LaTeXML to support the Digital Library of Mathematical Functions for creating web pages from mathematics papers written in LaTeX source files Miller [2014]. It is located at http://dlmf.nist.gov/LaTeXML/. Another large contributor to the project is the KWARC (Knowledge Adaptation and Reasoning for Content) group for bringing mathematics to the influence of the Semantic Web.

Usage: **latexml** *options* –destination=*doc*.xml *doc*

Bindings

LaTeXML bindings handle the translation of LaTeX style files into instructions that `latexml` can render into XML. Many bindings have already been written for the style files in wider use. If the binding hasn't been created for a paper which requires it, the `--includestyles` directive tells `latexml` to read any style files in its path and to proceed accordingly. In cases where `latexml` still fails to process the LaTeX file, a binding needs to be created. As we are dealing with LaTeX files crafted by over 1000 authors who are not always proficient in LaTeX (usually modifying someone else's document to fit their purpose), the quality of LaTeX varies from document to document. Rather than find the source of an error in LaTeX, authors can be tempted to add in any LaTeX command that will silence the warnings. Unfortunately, errors that the LaTeX processor will ignore can halt LaTeXML. One example is arXiv:hep-ex/0311034 where the author has used the `\abstract{}` command to start the paper which has a `\begin{indent}` directive, but not used a `maketitlepage` directive that calls the `\end{indent}` to close the abstract. To make LaTeX successfully parse the file, the author has added a `\end{indent}` at the end of the document. While this technique is efficient in author time, it leaves me in the position of having to correct the source

file if a suitable binding cannot be created.

XML

As mentioned in Section 4.2, the advantage of converting the document into an XML representation is the different parsing options available. Stream or event-based parsers are fast and memory efficient because they only examine one chunk of data at a time, process it and move on. This is fine if working with individual objects within the document is desired. When dealing with the document structure is preferred, a tree-based parser will examine the whole document and produce a hierarchical structure held in memory with which the document can be traversed very quickly. Generating that hierarchical structure is very costly in processor time, so it is not ideal for simple parsing tasks.

B.1 LaTeXML and Equations

LaTeXML can produce three XML representations of equations marked up in LaTeX format. The command `latexmlpost` can produce output in MathImages, MathML and OpenMath formats. These representations capture the mathematical semantics, easing the job of handling equations programatically.

C Algorithms and Software

This appendix contains a pseudocode description of the k-means and Louvain algorithms and issues relating to the software and database schema used to download and process the data.

k-means Algorithm

The k-means algorithm was first published in 1955 and is still widely used [Jain, 2010, Jain and Grewal, 2016]. It can be sensitive to the choice of initial centroid seeds, getting stuck in a local minimum and not finding the global minimum that produces the most optimal clusters. To avoid sub-optimal clusters, the algorithm should choose initial cluster centroids at random and be repeated several times to find the best set of clusters, as measured by the residual sum of squares of cluster members from their closest centroid.

```
Input: a list of documents, @Corpus
Output: the residual sum of squares (a measure of quality),
    a set of k clusters, @Clusters, each containing a document list

set initial cluster centroids, @Centroids = {C1 .. Ck},
    from k randomly selected documents from Corpus

while ($Residual_Sum_of_Squares > termination_condition) {
    # find closest centroid to paper
    foreach paper (@Corpus) {
        $max_similarity = 0;
        for ($i = 1; $i <= k; $i++) {
            $dot = dot_product(paper, $Centroid[i]);
            if ($dot > $max_similarity) {
                $max_similarity = $dot;
                $best_cluster   = $i;
            }
        }
        push $Clusters[best_cluster], paper;
    }
    # re-calculate cluster centroids
```

```
    for ($i = 1; $i <= k; $i++) {
        foreach paper ( $Clusters[$i] ) {
            cluster_terms += paper_terms;
        }
        centroid = sort_by_term_frequency(cluster_terms);
        truncate_to_first_1000_terms(centroid);
        $Centroids[$i] = centroid;
    }
    # calculate the Residual Sum of Squares
    for ($i = 1; $i <= k; $i++) {
        foreach paper ( $Clusters[$i] ) {
            $Residual_Sum_of_Squares += distance(paper, $Centroids[$i]);
        }
    }
}

return $Residual_Sum_of_Squares, @Clusters;
```

Louvain Community Detection Algorithm

Given a network of nodes, the algorithm finds the partition that maximises the modularity parameter, Q, given in Equation 3.2. The algorithm is not exact. It takes advantage of a property of the change in modularity, dQ given in Equation 3.3, to make the algorithm more efficient. It starts by assigning each node to its own community and then building larger communities by evaluating which communities to merge to increase modularity.

```
Input: a graph of nodes and links
Output: a file containing a list of nodes with their communities
        for each pass

foreach node i {
    place node i in community i
}

PASS: do {
    if (first iteration)
        Q_old = 0
    else
```

```
        Q_old = Q


    foreach node i {
        foreach neighbour j of node i {
            calculate dQ of removing i from its community
                and putting it with j
        }
        select community k with highest value of dQ
        if ( dQ > 0 )
            place node i in community k

    }
    calculate Q
} while ( Q > Q_old )

if (number of communities > 1) {
    store list of nodes with their communities
    foreach community j {
        new_link jk = sum (links from nodes in community j
            to nodes in community k)
    }
    foreach community i {
        new_node i = community i
    }
    # network now consists of new_nodes connected by new_links
    do PASS
}
```

Astro::ADS

`Astro::ADS` is a suite of four Perl modules written by Alastair Allen in 2002 as a part of the eSTAR project at University of Exeter. This bundle of modules is available via CPAN, the Comprehensive Perl Archive Network [Tregar, 2002].

Issues Raised by the Ampersand in Bibcodes

Because the ampersand, &, is an allowed character in ADS bibcodes, a conflict can arise with encoding the bibcode in the URL used to fetch a paper from the ADS

webservice.

The bibcode is constructed using codes derived from the year of publication, the journal in which is was published, its page and first author. There are a small number of journals in bibcode format that contain an ampersand, such as Astronomy & Astrophysics (A&A) and Annual Reviews of Astronomy & Astrophysics (ARA&A). When passed to a webserver unencoded, an ampersand is interpreted as the start of a new key-value pair to be handled by the server. This results in the server seeing a truncated bibcode and returning results from a different publication. In a large data gathering exercise, such erroneous results could introduce hard to explain features.

Missing References

When reporting the references received from the ADS, the loop stopped at the penultimate reference and exited leaving the last reference unreported. With the average paper having 20 references or more, this translates to a less than 5% error in the number of references reported.

Patches

The two patches submitted to the `Astro::ADS` author to rectify the ampersand bug and the missing references are listed below.

```
@@ -730,7 +732,10 @@

    # loop round all the options keys and build the query
    foreach my $key ( keys %{$self->{OPTIONS}} ) {
-       $options = $options . "&$key=${$self->{OPTIONS}}{$key}";
+       # some bibcodes have & and needs to be made "web safe"
+       my $websafe_option = ${$self->{OPTIONS}}{$key};
+       $websafe_option =~ s/&/%26/g;
+       $options = $options . "&$key=$websafe_option";
    }

    # build final query URL
@@ -903,8 +908,8 @@
```

```
          # LOOP THROUGH PAPER
          my ( @title, @authors, @affil, @journal, @pubdate, @keywords,
               @origin, @links, @url, @object, @abstract, @score );
-         while ( substr( $buffer[$counter], 0, 2 ) ne "%R" &&
-                 $counter < $#buffer ) {
+         while ( $counter <= $#buffer &&
+        substr( $buffer[$counter], 0, 2 ) ne "%R" ) {

             # grab the tags
             if( substr( $buffer[$counter], 0, 1 ) eq "%" ) {
```

MySQL Database Schema

The following statements are the commands to create the MySQL tables which describe the schema of the database used to store the collected data from the ADS.

```
create table ads_references ( paper_bib VARCHAR(19),
                              reference_bib VARCHAR(19) );
create table ads_citations ( paper_bib VARCHAR(19),
                             citation_bib VARCHAR(19) );
create table ads_data ( paper_bib VARCHAR(19) primary key,
            title VARCHAR(120), authors VARCHAR(350),
            keywords VARCHAR(256), journal VARCHAR(100),
            published VARCHAR(7) );
create table bibcodes_wanted ( bibcode varchar(19) );
create table bibcodes_fetched ( bibcode varchar(19) );

create table arxiv2bibcode ( arxiv_id VARCHAR(30),
                             bibcode VARCHAR(19),
                             category VARCHAR(18) );
create table bad_papers ( bibcode varchar(19),
                          reason varchar(40) );
create table citations_removed ( paper_bib VARCHAR(19),
                                 citation_bib VARCHAR(19),
                                 reason varchar(40) );
create table references_removed ( paper_bib VARCHAR(19),
                                  reference_bib VARCHAR(19),
                                  reason varchar(40) );
```

CLAIRLIB

The University of Michigan Computational Linguistics and Information Retrieval group, under the direction of Dragomir Radev, has produced the Clair Library suite of Perl modules to aid tasks in Natural Language Processing, Information Retrieval and Network Analysis [Radev et al., 2007]. Its most recent version, 1.08, was last updated September 2009.

D Mathematical Notes

This section contains the mathematical derivation of the error estimation due to incomplete normalisation of documents during document parsing that is presented in Section 4.3.3.

Incomplete Normalisation

To estimate the difference in cosine similarity as a result of the incomplete normalisation of a document, begin by calculating the error induced by one word having two forms, i.e. the distance between the two vectors that differ by the final term being reduced to create a new term.

Let us take the case of a generalised vector representation of a document where one term has two semantically equivalent forms.

Let the normalised vector have n terms with term-frequencies f_i and take the un-normalised representation to be the nth term split between two forms with frequencies, $f_n - g$ and g.

The cosine similarity between two documents is defined as the dot product of the vector representation of each document divided by the magnitudes of the vector representations.

$$sim(d_1, d_1') = \frac{\vec{V}(d_1) \cdot \vec{V}(d_1')}{\mid \vec{V}(d_1) \mid\mid \vec{V}(d_1') \mid}$$

where $\vec{V}(d_1)$ is the vector representation of the normalised document, d_1, and $\vec{V}(d_1')$ is the vector representation of the un-normalised document, d_1'.

Therefore, the cosine similarity between the normalised and un-normalised vectors is given by:

$$
\begin{aligned}
sim(d_1, d_1') &= \frac{\sum_{i=1}^{n-1} f_i^2 \; + f_n(f_n - g) + 0 \cdot g}{\sqrt{\sum_{i=1}^{n-1} f_i^2 \; + f_n^2}\sqrt{\sum_{i=1}^{n-1} f_i^2 \; + (f_n - g)^2 + g^2}} \\
&= \frac{\sum_{i=1}^{n} f_i^2 \; - f_n g}{\sqrt{\sum_{i=1}^{n} f_i^2}\sqrt{\sum_{i=1}^{n} f_i^2 + 2g^2 - 2f_n g}} \\
&= \frac{\sum_{i=1}^{n} f_i^2 \; - f_n g}{\sqrt{\left(\sum_{i=1}^{n} f_i^2\right)\left(\sum_{i=1}^{n} f_i^2 \; + 2g^2 - 2f_n g\right)}} \\
&= \sqrt{\frac{\left(\sum_{i=1}^{n} f_i^2 \; - f_n g\right)^2}{\left(\sum_{i=1}^{n} f_i^2\right)^2 + 2g^2 \sum_{i=1}^{n} f_i^2 \; - 2f_n g \sum_{i=1}^{n} f_i^2}} \\
&= \sqrt{\frac{\left(\sum_{i=1}^{n} f_i^2 \; - f_n g\right)^2}{\left(\sum_{i=1}^{n} f_i^2 - f_n g\right)^2 + 2g^2 \sum_{i=1}^{n} f_i^2 - g^2 f_n^2}} \\
&= \frac{1}{\sqrt{1 + \frac{2g^2 \sum_{i=1}^{n} f_i^2 - g^2 f_n^2}{(\sum_{i=1}^{n} f_i^2 \; - f_n g)^2}}}
\end{aligned}
$$

Two identical documents produce a maximum similarity of 1, therefore when we subtract this expression from 1, we can then produce a Maclaurin series (a power series expansion about the 0 point) to estimate the error produced by a single un-normalised

term. The form of the equation is

$$f(x) = 1 - \frac{1}{\sqrt{1+x}}$$

for which the Maclaurin series [Boas, 1983]

$$f(x) = f(0) + xf'(0) + \frac{1}{2!}x^2 f''(0) + \cdots + \frac{1}{n!}x^n f^{(n)}(0) + \cdots$$

is

$$f(x) = 0 + x \cdot \frac{1}{2}(1+x)^{-3/2} + \frac{x^2}{2} \cdot \frac{-3}{4}(1+x)^{-5/2} + \ldots$$

$$f(x) = \frac{1}{2}\sqrt{\frac{x^2}{(1+x)^3}} - \frac{3}{8}\sqrt{\frac{x^4}{(1+x)^5}} + \ldots$$

where $x = \frac{2g^2\sum_{i=1}^{n} f_i^2 - g^2 f_n^2}{(\sum_{i=1}^{n} f_i^2 - f_n g)^2}$

Ignoring the higher order terms which converge to zero because this is an alternating series with each higher term smaller than the preceding one, we see that

$$f(x) \approx x \cdot \frac{1}{2}(1+x)^{-3/2} < \frac{x}{2}$$

as $(1+x)^{-3/2} < 1$ with x positive.

Therefore, the error induced by a single un-normalised term is less than the following term.

$$error < \frac{2g^2\sum_{i=1}^{n} f_i^2 - g^2 f_n^2}{2(\sum_{i=1}^{n} f_i^2 - f_n g)^2} < \frac{g^2\sum_{i=1}^{n} f_i^2}{(\sum_{i=1}^{n} f_i^2 - f_n g)^2}$$

To give a simple example of that, let us take a small document with 50 terms found four times each in the document. Split one of the terms equally in two in order to produce the greatest effect. This gives us $n = 50$, $f_i = 4$ for all i and $g = 2$. For this small document, the error is less than 0.51%.

For future work to improve the utility of this calculation for estimating the effect of polysemy in Text Mining, consider the case of n unnormalised terms on the cosine similarity function and observe at what point the error overwhelms the similarity.

E The Dark Matter Problem

This appendix outlines an extremely short introduction to the Dark Matter problem to assist in understanding the keywords and labels encountered. A selected list for further reading on the topic can be found at the end.

Initial Observations

In astronomy, the information collected about the universe comes to us almost exclusively in the form of electronic radiation—light at all wavelengths. Estimates of mass are derived either from measuring the luminosity of an object and using empirical relationships between mass and luminosity, Υ, for different types of object or from its gravitational effects on nearby objects [Binney and Tremaine, 2008]. The motion of an object can be determined from its redshift, where a known wavelength of light is shifted towards longer wavelengths (a reddening) for motion away from us and shorter wavelengths for motion towards us (the equivalent of the Doppler shift for light).

The first evidence of Dark Matter was reported in Zwicky [1933] where the gravitational pull of several galaxies on other members of their galaxy cluster was found to exceed the mass determined by the mass derived from their visible light. The study of Dark Matter has increased exponentially since a clear example of this discrepancy was observed in 1969 by Vera Rubin and Kent Ford [Rubin and Ford, 1970]. They measured the rotation curves of spiral galaxies and found that the outer arms of the spirals were orbiting the galactic core faster than indicated by the luminosity profile.

Methods of Detection

Methods of detecting Dark Matter observationally range from measuring rotation curves of spiral galaxies to "gravitational lensing", using the bending of light by mass as

explained by Einstein's Theory of General Relativity [Blandford and Narayan, 1992]. COBE was a satellite experiment to measure fluctuations in the Cosmic Microwave Background, the afterglow of the Big Bang, and its observations were used to indicate which of the many models of the universe, frequently incorporating some type of Dark Matter such as Cold Dark Matter were more plausible than others [Ostriker, 1993].

Candidates

Candidates for the physical manifestation of Dark Matter are split into two classes, baryonic and non-baryonic [Feng, 2010].

Baryonic candidates are made of normal matter and include high mass objects (e.g. black holes, neutron stars and white dwarfs) and lower mass objects (e.g. brown dwarfs, planets, interstellar gas and dust) [Carr, 1994].

Non-Baryonic candidates are of interest to High Energy and Particle Physics community who are eager to try to explain the phenomena with exotic particles or new forms of matter and seek out the existence and character of particles such as the neutrino, ν, the gravitino, $\bar{G}$ and others in the class called Weakly Interacting Massive Particles (WIMP). One theoretical particle of great interest is the sterile neutrino which only interacts via gravity and not the weak force. Evidence (or the lack of it) for these particles is found in neutrino detectors (such as the Sudbury Neutrino Observatory) and particle colliders (such as the Large Hadron Collider).

Alternative Theories

Until Dark Matter is detected directly, there will always be some researchers exploring the hypothesis that Dark Matter does *not* exist and that our understanding of gravity at large scales requires re-evaluation [Sanders and McGaugh, 2002]. Very much a minority position, MOND (MOdified Newtonian Dynamics) theorists propose a model of new

physics that alters the equations of Newton and Einstein at distances greater than $10^{18}m$ to account for discrepancies between observations and standard gravitational theory.

Other alternatives include Brane cosmology. Related to string theory, it proposes that dark matter is a manifestation of matter existing in higher dimensions than our normal three dimensions of space. Attempting to explain the Standard Model of physics, it bridges the fields of cosmology and particle physics.

The Bullet Cluster - a Test of Two Theories

A test of the two competing theories, Dark Matter and MOND, was found in the Bullet Cluster (X-ray catalogue identifier 1E 0657-558), a pair of galaxies which have collided and were discovered in 2006. Observations seemed to favour the Dark Matter models over MOND, but the interpretation is still disputed.

References

An excellent introduction to cosmology is *Big Bang* by Silk [2001]. It uses no mathematics in the main portion of the text, yet manages to explain the phenomena at a deep level. General introductions to Dark Matter range from *Structure of the Universe* by Gribbin [1990] to *Quintessence* by Krauss [2001], but the only mention of MOND is a few pages in *Origins* by Tyson [2004]. *Galactic Dynamics* by Binney and Tremaine [2008] is an advanced undergraduate textbook also used in post-graduate studies. Researchers seeking an introduction to the topic are directed to several reviews in the Annual Reviews of Astronomy & Astrophysics by the following: Trimble [1987], Carr [1994], Sanders and McGaugh [2002], Ostriker [1993], Porter et al. [2011], Blandford and Narayan [1992] and Feng [2010].

F Stop Words and Group Data

Lingua::EN::StopWords

This list of stopwords is used by the Perl module `Lingua::EN::StopWords` to assist in removing words before processing.

Table F.1: Lingua::EN::StopWords

these	forth	which	both	another	nobody	many
far	if	himself	him	own	always	also
deep	selves	each	what	although	them	inward
your	somebody	but	again	too	and	several
over	of	kept	still	is	all	being
she	will	nor	have	much	neither	when
said	it	nowhere	every	either	can	besides
where	instead	gets	within	often	were	next
yourself	a	would	almost	off	no	thus
etc	might	got	in	quite	upon	very
somewhat	ours	only	toward	anything	by	myself
mine	they	whose	after	doing	am	his
get	i	anyone	under	thoroughly	across	well
through	there	while	adj	as	anywhere	themselves
please	everybody	nothing	everyone	because	itself	however
around	others	along	cannot	has	that	not
whatever	on	our	who	shall	mostly	its
out	whenever	theirs	ought	seem	some	hardly
with	apart	behind	here	must	none	aside

Continued on next page

did	do	into	to	herself	indeed	from
towards	really	her	below	during	any	pp
plus	alone	then	downwards	more	an	the
against	done	onto	yet	since	self	or
could	does	anybody	few	down	about	whom
this	before	so	enough	for	how	v
per	outside	else	among	near	those	rather
their	other	be	whether	such	away	thorough
ever	even	most	therefore	young	up	beyond
are	having	been	above	should	at	together
except	had	maybe	between	without	than	until
p	was	just				

corpus tf pruned

These words were found in top 200 most common words in the corpus and are included in the extended Stop List.

Table F.2: Extended stopword list derived from the Dark Matter Corpus

we	one	s	two	al	et	e	given
results	see	fig	value	number	may	section	using
high	values	low	eq	g	first	figure	range
same	non	used	due	present	form	shown	larger
found	show	find	use	possible	lower	expected	smaller

Group Data

Table F.3: Constituent Data for Group Membership in the citation network clustering. The Group id is an arbitrary number assigned by the Louvain clustering program. N is the number of members in the group. Columns **a, h, m** and **o** refer to the percentage of N members which belong to categories `astro-ph, hep, mond` and `other`. Label is made from the top seven keywords given to the members of the group.

Group id	N	a	h	m	o	Label
92	476	72	28	0	1	high energy physics; phenomenology; supersymmetric models; galaxy; theory; experiment; supersymmetry
211	45	100	0	0	0	gravitational lensing; theory; halos; quasars; general; formation; large-scale structure of universe
213	642	99	0	0	0	theory; methods; clusters; formation; large-scale structure of universe; general; halos
263	3	100	0	0	0	phenomenology; high energy physics
377	2	0	0	0	100	general physics
424	10	100	0	0	0	black hole physics; nuclei; population iii stars; spatial distribution of galaxies; galactic halo; structure; evolution
486	24	83	17	0	0	high energy physics; phenomenology; extensions of electroweak gauge sector; gamma-ray sources; other gauge bosons; unidentified sources of radiation outside the solar system; radio
544	52	88	12	0	0	x-rays; individual; neutrinos; particle-theory and field-theory models of the early universe; black hole physics; elementary particles; galaxy
587	2	100	0	0	0	technicolor models; phenomenology; high energy physics; particle-theory and field-theory models of the early universe
683	2	0	0	0	100	particle-theory and field-theory models of the early universe; cosmic rays
691	109	24	1	67	8	gravitation; kinematics and dynamics; clusters; theory; general; high energy physics; general relativity and quantum cosmology
700	56	93	0	5	2	kinematics and dynamics; structure; individual; halos; ism; evolution; spiral
719	2	100	0	0	0	dwarf; irregular; evolution; abundances
726	2	100	0	0	0	galaxy; clouds; submillimetre; evolution; halo; starburst; large-scale structure of universe
739	255	72	9	1	18	theory; high energy physics; dark energy; general relativity and quantum cosmology; particle-theory and field-theory models of the early universe; phenomenology; observational cosmology
808	2	0	100	0	0	high energy physics; phenomenology; theory
811	13	100	0	0	0	gravitational lensing; large-scale structure of universe; methods; image processing; theory; numerical; astronomical techniques
900	5	100	0	0	0	neutrinos; phenomenology; muons; neutrino oscillations; neutrino mass; nonluminous matter; supernovae
914	73	99	0	0	1	galaxy; gravitational lensing; halo; stars; gravitational microlensing; ism; brown dwarfs
949	7	100	0	0	0	phenomenology; high energy physics; galaxy; black hole physics; center; hypothetical particles; stellar evolution
1034	2	50	0	0	50	hubble law; carmeli cosmology; anomalous acceleration; mond
1078	25	96	0	4	0	cosmic microwave background; clusters; general; theory; large-scale structure of universe; intergalactic medium; gravitational lensing
1082	20	95	0	0	5	cosmic rays; hypothetical particles; phenomenology; early universe; superheavy dark matter; neutrino; active and peculiar galaxies and related systems

Continued on next page

Group id	N	a	h	m	o	Label
1125	63	94	0	6	0	kinematics and dynamics; galaxy; methods; dwarf; local group; structure; evolution
1165	2	100	0	0	0	phenomenology; high energy physics
1281	2	100	0	0	0	supersymmetry; gamma-ray sources; gamma-ray bursts
1391	3	100	0	0	0	protoplanetary disks; planetary systems; instabilities; stars; magnetic fields; mhd; magnetohydrodynamics
1400	5	100	0	0	0	theory; bursts; high energy physics; gamma rays; galaxy; phenomenology; early universe
1552	6	83	0	17	0	star clusters; individual; clusters; general; messier number; stellar dynamics; stellar content
1705	22	73	14	0	14	theory; cosmic strings; high energy physics; particle-theory and field-theory models of the early universe; phenomenology; large-scale structure of universe; spatial distribution of galaxies
1814	44	98	0	2	0	elliptical and lenticular; kinematics and dynamics; cd; individual; halos; haloes; structure
1848	2	100	0	0	0	high energy physics; theory; phenomenology; general relativity and quantum cosmology
4585	8	100	0	0	0	galaxy; stars; white dwarfs; stellar content; structure; halo; luminosity function
4850	55	100	0	0	0	theory; formation; stars; early universe; evolution; intergalactic medium; quasars
4910	11	9	91	0	0	neutrino mass and mixing; neutrino mass; baryogenesis; phenomenology; quark and lepton masses and mixing; black holes; primordial galaxies
5881	6	100	0	0	0	baryogenesis; composition; detection; neutrinos; nucleosynthesis
5930	15	87	13	0	0	phenomenology; high energy physics; cosmic microwave background; big bang nucleosynthesis; large scale structure; elementary particle processes; galaxy c
5931	40	42	55	0	2	high energy physics; phenomenology; particle-theory and field-theory models of the early universe; supersymmetric models; early universe; physics; experiment
5940	13	100	0	0	0	formation; theory; large-scale structure of universe; evolution; stellar content; methods; halos
5955	11	27	45	0	27	theory; high energy physics; general relativity and quantum cosmology; extensions of electroweak gauge sector; brane; models beyond the standard model; field theories in dimensions other than four
5960	5	100	0	0	0	ism; clouds; large-scale structure of universe; molecules
5966	5	100	0	0	0	galaxy; ism; structure; gamma rays; bubbles; local group; clouds
5967	13	85	15	0	0	axions and other nambu-goldstone bosons; galaxy; dwarf; instrumentation; radio, microwave; particle emission; solar neutrinos
5969	3	100	0	0	0	shock waves; instabilities; radio; x-ray; microwave; gamma-ray; acceleration of particles
5973	77	52	48	0	0	cosmic rays; high energy physics; phenomenology; supersymmetric partners of known particles; supersymmetric models; galactic halo; gamma-ray

Group Similarity measurements

Table F.4: Group comparisons of Similarity, Group Size and Link Saturation for 31 citation network communities presented in Figure 5.3 (2 column layout).

Group ids		$\overline{\text{Similarity}}$	σ	N_1	N_2	Links	**Group ids**		$\overline{\text{Similarity}}$	σ	N_1	N_2	Links
92	92	0.315	0.128	476	476	5182	92	211	0.212	0.080	476	45	13
92	213	0.233	0.097	476	642	621	92	263	0.277	0.083	476	3	11
92	321	0.215	0.081	476	2	0	92	377	0.252	0.080	476	2	0
92	486	0.325	0.113	476	24	100	92	544	0.261	0.098	476	52	53
92	587	0.356	0.097	476	2	9	92	683	0.271	0.092	476	2	0
92	700	0.220	0.092	476	56	18	92	719	0.203	0.077	476	2	0
92	726	0.179	0.063	476	2	0	92	739	0.254	0.092	476	255	78
92	808	0.272	0.094	476	2	1	92	822	0.196	0.075	476	2	2
92	900	0.253	0.111	476	5	3	92	914	0.218	0.085	476	73	22
92	949	0.257	0.089	476	7	6	92	957	0.197	0.079	476	2	0
92	4585	0.145	0.055	476	8	0	92	4850	0.241	0.101	476	55	83
92	4903	0.233	0.084	476	2	3	92	5092	0.241	0.090	476	3	5
92	5881	0.266	0.122	476	6	0	92	5931	0.250	0.107	476	40	54
92	5957	0.222	0.067	476	2	1	92	5960	0.207	0.089	476	5	0
92	5966	0.243	0.095	476	5	3	92	5969	0.254	0.107	476	3	9
92	5973	0.308	0.110	476	77	383	211	211	0.401	0.120	45	45	94
211	213	0.299	0.100	45	642	116	211	263	0.190	0.044	45	3	0
211	321	0.210	0.050	45	2	0	211	377	0.259	0.055	45	2	0
211	486	0.227	0.072	45	24	0	211	544	0.227	0.080	45	52	9
211	587	0.236	0.048	45	2	0	211	683	0.161	0.048	45	2	0
211	700	0.292	0.097	45	56	4	211	719	0.261	0.072	45	2	0
211	726	0.273	0.072	45	2	0	211	739	0.236	0.081	45	255	8
211	808	0.292	0.061	45	2	0	211	822	0.174	0.052	45	2	0
211	900	0.191	0.080	45	5	0	211	914	0.276	0.089	45	73	9
211	949	0.227	0.066	45	7	0	211	957	0.254	0.081	45	2	1
211	4585	0.184	0.056	45	8	0	211	4850	0.223	0.086	45	55	1
211	4903	0.163	0.044	45	2	0	211	5092	0.178	0.073	45	3	0
211	5881	0.307	0.085	45	6	0	211	5931	0.193	0.080	45	40	0
211	5957	0.148	0.034	45	2	0	211	5960	0.204	0.051	45	5	0
211	5966	0.300	0.085	45	5	0	211	5969	0.198	0.052	45	3	0
211	5973	0.214	0.068	45	77	2	213	213	0.349	0.116	642	642	7232
213	263	0.214	0.074	642	3	3	213	321	0.259	0.071	642	2	0
213	377	0.275	0.072	642	2	0	213	486	0.245	0.096	642	24	22
213	544	0.255	0.099	642	52	214	213	587	0.267	0.080	642	2	1
213	683	0.216	0.092	642	2	0	213	700	0.311	0.110	642	56	178
213	719	0.260	0.079	642	2	0	213	726	0.257	0.091	642	2	2
213	739	0.260	0.096	642	255	205	213	808	0.302	0.089	642	2	1
213	822	0.232	0.074	642	2	0	213	900	0.202	0.099	642	5	7
213	914	0.236	0.095	642	73	19	213	949	0.239	0.083	642	7	1
213	957	0.266	0.070	642	2	0	213	4585	0.180	0.053	642	8	1
213	4850	0.254	0.098	642	55	94	213	4903	0.201	0.078	642	2	10
213	5092	0.196	0.081	642	3	0	213	5881	0.313	0.127	642	6	0
213	5931	0.211	0.096	642	40	50	213	5957	0.179	0.058	642	2	0
213	5960	0.223	0.072	642	5	0	213	5966	0.293	0.083	642	5	2
213	5969	0.241	0.077	642	3	0	213	5973	0.230	0.083	642	77	83
263	263	0.773	0.090	3	3	6	263	321	0.244	0.081	3	2	0
263	377	0.292	0.031	3	2	0	263	486	0.330	0.069	3	24	0
263	544	0.271	0.060	3	52	0	263	587	0.605	0.165	3	2	3
263	683	0.419	0.064	3	2	0	263	700	0.198	0.061	3	56	0

Continued on next page

Group ids		$\overline{\text{Similarity}}$	σ	N_1	N_2	Links	Group ids		$\overline{\text{Similarity}}$	σ	N_1	N_2	Links
263	719	0.202	0.073	3	2	0	263	726	0.143	0.033	3	2	0
263	739	0.290	0.068	3	255	0	263	808	0.322	0.040	3	2	0
263	822	0.155	0.027	3	2	0	263	900	0.259	0.073	3	5	0
263	914	0.198	0.077	3	73	0	263	949	0.254	0.044	3	7	0
263	957	0.214	0.030	3	2	0	263	4585	0.122	0.035	3	8	0
263	4850	0.232	0.067	3	55	2	263	4903	0.344	0.071	3	2	0
263	5092	0.292	0.036	3	3	0	263	5881	0.241	0.101	3	6	0
263	5931	0.318	0.123	3	40	3	263	5957	0.262	0.021	3	2	0
263	5960	0.218	0.090	3	5	0	263	5966	0.218	0.055	3	5	0
263	5969	0.237	0.030	3	3	0	263	5973	0.310	0.071	3	77	1
321	377	0.435	0.039	2	2	0	321	486	0.227	0.067	2	24	0
321	544	0.218	0.064	2	52	0	321	587	0.290	0.070	2	2	0
321	683	0.264	0.083	2	2	0	321	700	0.284	0.083	2	56	0
321	719	0.195	0.014	2	2	0	321	726	0.175	0.052	2	2	0
321	739	0.332	0.097	2	255	0	321	808	0.263	0.062	2	2	0
321	822	0.243	0.046	2	2	0	321	900	0.151	0.060	2	5	0
321	914	0.204	0.067	2	73	0	321	949	0.262	0.088	2	7	0
321	957	0.263	0.062	2	2	0	321	4585	0.152	0.044	2	8	0
321	4850	0.202	0.065	2	55	0	321	4903	0.219	0.080	2	2	0
321	5092	0.204	0.063	2	3	0	321	5881	0.247	0.080	2	6	0
321	5931	0.202	0.075	2	40	0	321	5957	0.200	0.048	2	2	0
321	5960	0.266	0.106	2	5	0	321	5966	0.248	0.054	2	5	0
321	5969	0.221	0.046	2	3	0	321	5973	0.210	0.058	2	77	0
377	486	0.267	0.066	2	24	0	377	544	0.255	0.067	2	52	0
377	587	0.372	0.040	2	2	0	377	683	0.269	0.023	2	2	0
377	700	0.305	0.073	2	56	2	377	719	0.239	0.024	2	2	0
377	726	0.229	0.016	2	2	0	377	739	0.347	0.092	2	255	0
377	808	0.369	0.044	2	2	0	377	822	0.254	0.049	2	2	0
377	900	0.237	0.086	2	5	0	377	914	0.263	0.078	2	73	0
377	949	0.283	0.053	2	7	0	377	957	0.359	0.029	2	2	0
377	4585	0.173	0.047	2	8	0	377	4850	0.210	0.067	2	55	0
377	4903	0.255	0.052	2	2	0	377	5092	0.232	0.027	2	3	0
377	5881	0.333	0.084	2	6	0	377	5931	0.232	0.086	2	40	0
377	5957	0.213	0.021	2	2	0	377	5960	0.271	0.065	2	5	0
377	5966	0.293	0.054	2	5	0	377	5969	0.225	0.032	2	3	0
377	5973	0.236	0.058	2	77	0	486	486	0.482	0.120	24	24	154
486	544	0.280	0.087	24	52	7	486	587	0.397	0.082	24	2	0
486	683	0.360	0.088	24	2	0	486	700	0.237	0.081	24	56	0
486	719	0.239	0.081	24	2	0	486	726	0.201	0.067	24	2	0
486	739	0.275	0.086	24	255	4	486	808	0.314	0.084	24	2	0
486	822	0.205	0.063	24	2	0	486	900	0.239	0.086	24	5	0
486	914	0.234	0.083	24	73	0	486	949	0.292	0.074	24	7	0
486	957	0.206	0.052	24	2	0	486	4585	0.161	0.062	24	8	0
486	4850	0.271	0.108	24	55	17	486	4903	0.262	0.084	24	2	1
486	5092	0.249	0.072	24	3	0	486	5881	0.286	0.105	24	6	0
486	5931	0.268	0.103	24	40	2	486	5957	0.253	0.051	24	2	0
486	5960	0.215	0.081	24	5	0	486	5966	0.276	0.092	24	5	0
486	5969	0.302	0.118	24	3	0	486	5973	0.382	0.109	24	77	22
544	544	0.443	0.154	52	52	902	544	587	0.361	0.092	52	2	0
544	683	0.279	0.078	52	2	0	544	700	0.231	0.086	52	56	1
544	719	0.218	0.078	52	2	0	544	726	0.202	0.074	52	2	0
544	739	0.266	0.087	52	255	17	544	808	0.300	0.073	52	2	0
544	822	0.182	0.057	52	2	0	544	900	0.420	0.154	52	5	0
544	914	0.214	0.085	52	73	1	544	949	0.307	0.107	52	7	6

Continued on next page

Group ids		$\overline{\text{Similarity}}$	σ	N_1	N_2	Links	Group ids		$\overline{\text{Similarity}}$	σ	N_1	N_2	Links
544	957	0.237	0.072	52	2	0	544	4585	0.154	0.058	52	8	0
544	4850	0.254	0.102	52	55	55	544	4903	0.218	0.068	52	2	0
544	5092	0.229	0.073	52	3	0	544	5881	0.284	0.120	52	6	0
544	5931	0.269	0.105	52	40	24	544	5957	0.227	0.059	52	2	0
544	5960	0.235	0.102	52	5	0	544	5966	0.244	0.077	52	5	0
544	5969	0.238	0.087	52	3	0	544	5973	0.281	0.087	52	77	15
587	683	0.406	0.017	2	2	0	587	700	0.243	0.067	2	56	0
587	719	0.234	0.086	2	2	0	587	726	0.179	0.043	2	2	0
587	739	0.342	0.076	2	255	0	587	808	0.390	0.091	2	2	0
587	822	0.224	0.024	2	2	0	587	900	0.388	0.123	2	5	0
587	914	0.240	0.086	2	73	0	587	949	0.306	0.053	2	7	0
587	957	0.257	0.039	2	2	0	587	4585	0.157	0.038	2	8	0
587	4850	0.294	0.092	2	55	0	587	4903	0.357	0.089	2	2	0
587	5092	0.328	0.055	2	3	0	587	5881	0.308	0.132	2	6	0
587	5931	0.349	0.110	2	40	1	587	5957	0.318	0.023	2	2	0
587	5960	0.263	0.099	2	5	0	587	5966	0.278	0.058	2	5	0
587	5969	0.283	0.064	2	3	0	587	5973	0.379	0.081	2	77	0
683	700	0.201	0.072	2	56	0	683	719	0.174	0.055	2	2	0
683	726	0.131	0.020	2	2	0	683	739	0.279	0.091	2	255	0
683	808	0.314	0.045	2	2	0	683	822	0.177	0.025	2	2	0
683	900	0.199	0.068	2	5	0	683	914	0.181	0.081	2	73	0
683	949	0.303	0.076	2	7	0	683	957	0.194	0.014	2	2	0
683	4585	0.094	0.028	2	8	0	683	4850	0.215	0.089	2	55	0
683	4903	0.300	0.116	2	2	0	683	5092	0.256	0.016	2	3	0
683	5881	0.229	0.090	2	6	0	683	5931	0.286	0.108	2	40	0
683	5957	0.296	0.019	2	2	0	683	5960	0.211	0.111	2	5	0
683	5966	0.205	0.039	2	5	0	683	5969	0.262	0.013	2	3	0
683	5973	0.296	0.070	2	77	0	700	700	0.394	0.140	56	56	152
700	719	0.326	0.084	56	2	1	700	726	0.268	0.079	56	2	0
700	739	0.249	0.096	56	255	26	700	808	0.287	0.079	56	2	0
700	822	0.225	0.072	56	2	0	700	900	0.181	0.081	56	5	0
700	914	0.252	0.099	56	73	6	700	949	0.245	0.077	56	7	2
700	957	0.262	0.066	56	2	0	700	4585	0.193	0.063	56	8	0
700	4850	0.225	0.092	56	55	2	700	4903	0.181	0.059	56	2	0
700	5092	0.176	0.070	56	3	0	700	5881	0.353	0.134	56	6	4
700	5931	0.192	0.087	56	40	0	700	5957	0.161	0.048	56	2	0
700	5960	0.229	0.071	56	5	0	700	5966	0.333	0.104	56	5	4
700	5969	0.217	0.062	56	3	0	700	5973	0.219	0.077	56	77	0
719	726	0.276	0.054	2	2	0	719	739	0.229	0.088	2	255	0
719	808	0.278	0.079	2	2	0	719	822	0.166	0.019	2	2	0
719	900	0.186	0.080	2	5	0	719	914	0.246	0.087	2	73	0
719	949	0.233	0.057	2	7	0	719	957	0.206	0.024	2	2	0
719	4585	0.256	0.049	2	8	0	719	4850	0.228	0.083	2	55	0
719	4903	0.180	0.046	2	2	0	719	5092	0.161	0.081	2	3	0
719	5881	0.303	0.074	2	6	0	719	5931	0.187	0.088	2	40	0
719	5957	0.154	0.065	2	2	0	719	5960	0.205	0.055	2	5	0
719	5966	0.297	0.043	2	5	0	719	5969	0.209	0.058	2	3	0
719	5973	0.219	0.074	2	77	0	726	739	0.195	0.062	2	255	0
726	808	0.234	0.068	2	2	0	726	822	0.153	0.044	2	2	0
726	900	0.160	0.052	2	5	0	726	914	0.265	0.099	2	73	5
726	949	0.237	0.084	2	7	0	726	957	0.236	0.042	2	2	0
726	4585	0.224	0.068	2	8	0	726	4850	0.216	0.083	2	55	0
726	4903	0.157	0.035	2	2	0	726	5092	0.130	0.045	2	3	0
726	5881	0.259	0.085	2	6	0	726	5931	0.159	0.067	2	40	0

Continued on next page

Group ids		$\overline{\text{Similarity}}$	σ	N_1	N_2	Links	Group ids		$\overline{\text{Similarity}}$	σ	N_1	N_2	Links
726	5957	0.131	0.036	2	2	0	726	5960	0.228	0.114	2	5	0
726	5966	0.324	0.090	2	5	0	726	5969	0.213	0.087	2	3	0
726	5973	0.182	0.059	2	77	0	739	739	0.374	0.123	255	255	1260
739	808	0.349	0.086	255	2	0	739	822	0.186	0.051	255	2	0
739	900	0.238	0.099	255	5	0	739	914	0.218	0.090	255	73	16
739	949	0.247	0.072	255	7	0	739	957	0.275	0.071	255	2	0
739	4585	0.145	0.051	255	8	0	739	4850	0.237	0.090	255	55	4
739	4903	0.258	0.080	255	2	1	739	5092	0.271	0.084	255	3	0
739	5881	0.276	0.108	255	6	0	739	5931	0.258	0.098	255	40	4
739	5957	0.257	0.065	255	2	0	739	5960	0.256	0.093	255	5	0
739	5966	0.252	0.083	255	5	0	739	5969	0.249	0.064	255	3	0
739	5973	0.273	0.083	255	77	8	808	822	0.225	0.022	2	2	0
808	900	0.261	0.116	2	5	0	808	914	0.285	0.102	2	73	0
808	949	0.271	0.039	2	7	0	808	957	0.375	0.065	2	2	0
808	4585	0.180	0.054	2	8	0	808	4850	0.285	0.091	2	55	0
808	4903	0.336	0.020	2	2	0	808	5092	0.267	0.083	2	3	0
808	5881	0.389	0.137	2	6	0	808	5931	0.321	0.110	2	40	0
808	5957	0.268	0.020	2	2	0	808	5960	0.289	0.101	2	5	0
808	5966	0.321	0.044	2	5	0	808	5969	0.216	0.076	2	3	0
808	5973	0.279	0.071	2	77	0	822	900	0.122	0.042	2	5	0
822	914	0.162	0.059	2	73	0	822	949	0.218	0.047	2	7	0
822	957	0.234	0.035	2	2	0	822	4585	0.136	0.048	2	8	0
822	4850	0.211	0.058	2	55	0	822	4903	0.171	0.045	2	2	0
822	5092	0.118	0.043	2	3	0	822	5881	0.230	0.096	2	6	0
822	5931	0.156	0.054	2	40	0	822	5957	0.141	0.015	2	2	0
822	5960	0.155	0.032	2	5	0	822	5966	0.230	0.032	2	5	0
822	5969	0.180	0.040	2	3	0	822	5973	0.177	0.050	2	77	0
900	900	0.593	0.184	5	5	2	900	914	0.194	0.082	5	73	0
900	949	0.278	0.143	5	7	0	900	957	0.197	0.089	5	2	0
900	4585	0.122	0.047	5	8	0	900	4850	0.196	0.086	5	55	0
900	4903	0.191	0.066	5	2	0	900	5092	0.194	0.084	5	3	0
900	5881	0.259	0.160	5	6	0	900	5931	0.226	0.104	5	40	0
900	5957	0.182	0.053	5	2	0	900	5960	0.216	0.134	5	5	0
900	5966	0.201	0.085	5	5	0	900	5969	0.186	0.058	5	3	0
900	5973	0.249	0.084	5	77	0	914	914	0.358	0.145	73	73	186
914	949	0.241	0.076	73	7	0	914	957	0.259	0.103	73	2	1
914	4585	0.264	0.125	73	8	11	914	4850	0.209	0.085	73	55	2
914	4903	0.196	0.077	73	2	0	914	5092	0.162	0.075	73	3	0
914	5881	0.306	0.114	73	6	1	914	5931	0.195	0.100	73	40	1
914	5957	0.152	0.055	73	2	0	914	5960	0.219	0.097	73	5	2
914	5966	0.291	0.105	73	5	0	914	5969	0.188	0.076	73	3	0
914	5973	0.207	0.075	73	77	1	949	949	0.407	0.175	7	7	0
949	957	0.255	0.066	7	2	0	949	4585	0.179	0.061	7	8	0
949	4850	0.236	0.089	7	55	0	949	4903	0.225	0.051	7	2	0
949	5092	0.184	0.049	7	3	0	949	5881	0.276	0.075	7	6	0
949	5931	0.220	0.083	7	40	1	949	5957	0.189	0.040	7	2	0
949	5960	0.230	0.085	7	5	0	949	5966	0.267	0.079	7	5	0
949	5969	0.241	0.100	7	3	0	949	5973	0.259	0.068	7	77	0
957	4585	0.169	0.044	2	8	0	957	4850	0.235	0.077	2	55	0
957	4903	0.237	0.033	2	2	0	957	5092	0.196	0.027	2	3	0
957	5881	0.305	0.071	2	6	0	957	5931	0.220	0.084	2	40	0
957	5957	0.204	0.038	2	2	0	957	5960	0.272	0.050	2	5	0
957	5966	0.277	0.049	2	5	0	957	5969	0.207	0.039	2	3	0
957	5973	0.186	0.052	2	77	0	4585	4585	0.594	0.172	8	8	16

Continued on next page

Group ids		$\overline{\text{Similarity}}$	σ	N_1	N_2	Links	Group ids		$\overline{\text{Similarity}}$	σ	N_1	N_2	Links
4585	4850	0.163	0.058	8	55	0	4585	4903	0.121	0.028	8	2	0
4585	5092	0.099	0.042	8	3	0	4585	5881	0.194	0.056	8	6	0
4585	5931	0.130	0.063	8	40	0	4585	5957	0.100	0.024	8	2	0
4585	5960	0.145	0.044	8	5	0	4585	5966	0.212	0.057	8	5	0
4585	5969	0.122	0.040	8	3	0	4585	5973	0.141	0.045	8	77	0
4850	4850	0.333	0.150	55	55	256	4850	4903	0.201	0.068	55	2	0
4850	5092	0.182	0.073	55	3	0	4850	5881	0.250	0.096	55	6	0
4850	5931	0.225	0.093	55	40	3	4850	5957	0.203	0.062	55	2	0
4850	5960	0.219	0.087	55	5	0	4850	5966	0.236	0.080	55	5	0
4850	5969	0.248	0.107	55	3	0	4850	5973	0.256	0.103	55	77	11
4903	5092	0.229	0.066	2	3	0	4903	5881	0.243	0.101	2	6	0
4903	5931	0.242	0.093	2	40	1	4903	5957	0.230	0.052	2	2	0
4903	5960	0.203	0.072	2	5	0	4903	5966	0.209	0.047	2	5	0
4903	5969	0.187	0.056	2	3	0	4903	5973	0.231	0.076	2	77	1
5092	5092	0.526	0.172	3	3	0	5092	5881	0.215	0.098	3	6	0
5092	5931	0.261	0.084	3	40	0	5092	5957	0.253	0.037	3	2	0
5092	5960	0.180	0.068	3	5	0	5092	5966	0.204	0.089	3	5	0
5092	5969	0.181	0.049	3	3	0	5092	5973	0.267	0.084	3	77	0
5881	5881	0.377	0.189	6	6	0	5881	5931	0.236	0.123	6	40	0
5881	5957	0.190	0.065	6	2	0	5881	5960	0.248	0.101	6	5	0
5881	5966	0.351	0.111	6	5	0	5881	5969	0.229	0.096	6	3	0
5881	5973	0.258	0.096	6	77	0	5931	5931	0.352	0.150	40	40	70
5931	5957	0.277	0.096	40	2	0	5931	5960	0.204	0.097	40	5	0
5931	5966	0.209	0.090	40	5	0	5931	5969	0.206	0.075	40	3	0
5931	5973	0.279	0.107	40	77	23	5957	5960	0.178	0.050	2	5	0
5957	5966	0.184	0.042	2	5	0	5957	5969	0.187	0.025	2	3	0
5957	5973	0.260	0.065	2	77	4	5960	5960	0.265	0.179	5	5	2
5960	5966	0.250	0.070	5	5	0	5960	5969	0.219	0.079	5	3	0
5960	5973	0.210	0.079	5	77	0	5966	5966	0.395	0.112	5	5	0
5966	5969	0.203	0.067	5	3	0	5966	5973	0.245	0.086	5	77	3
5969	5969	0.334	0.109	3	3	0	5969	5973	0.284	0.112	3	77	3
5973	5973	0.372	0.121	77	77	422	Ensemble		0.259	0.109			

Text corpus from arXiv

The following are the 2661 arXiv identifiers that were used to construct the Dark Matter corpus.

Table F.5: arXiv ids constituting the Dark Matter corpus

physics/9911007	physics/9910044	physics/9908015	physics/9812021	physics/9808051	physics/0702132
physics/0702019	physics/0612072	physics/0611273	physics/0610136	physics/0610135	physics/0508098
physics/0506042	physics/0506002	physics/0504151	physics/0503079	physics/0408126	physics/0406159
physics/0405147	physics/0402075	physics/0401047	physics/0006040	nucl-th/9509026	nucl-th/0610120
nucl-th/0411021	nucl-ex/0702031	nucl-ex/0512012	nucl-ex/0410025	nucl-ex/0409014	nucl-ex/0202014
nucl-ex/0110003	hep-th/9610210	hep-th/0703070	hep-th/0702212	hep-th/0610240	hep-th/0602136
hep-th/0508161	hep-th/0507199	hep-th/0507182	hep-th/0505042	hep-th/0503062	hep-th/0411158
hep-th/0411025	hep-th/0404170	hep-th/0404099	hep-th/0403054	hep-th/0309150	hep-th/0307028
hep-th/0205207	hep-th/0205055	hep-th/0112036	hep-th/0110208	hep-th/0107259	hep-th/0103234
hep-th/0005033	hep-ph/0703310	hep-ph/0703181	hep-ph/0703130	hep-ph/0703097	hep-ph/0703056
hep-ph/0703024	hep-ph/0703014	hep-ph/0702223	hep-ph/0702184	hep-ph/0702143	hep-ph/0702080
hep-ph/0702051	hep-ph/0702041	hep-ph/0701271	hep-ph/0701266	hep-ph/0701233	hep-ph/0701229
hep-ph/0701197	hep-ex/9905042	hep-ex/9904034	hep-ex/9904005	hep-ex/9901021	hep-ex/9812020
hep-ex/9811040	hep-ex/9811022	hep-ex/9804007	hep-ex/9802007	hep-ex/9709019	hep-ex/9612014
hep-ex/0701025	hep-ex/0509004	hep-ex/0507086	hep-ex/0505053	hep-ex/0504031	hep-ex/0504022
hep-ex/0404042	hep-ex/0404025	hep-ex/0401032	hep-ex/0312049	hep-ex/0311034	hep-ex/0306001
hep-ex/0302022	hep-ex/0301039	hep-ex/0212055	hep-ex/0111073	hep-ex/0109013	hep-ex/0106015
hep-ex/0102013	hep-ex/0006015	hep-ex/0005031	hep-ex/0005003	gr-qc/9910048	gr-qc/9905101
gr-qc/9810028	gr-qc/9802054	gr-qc/9705024	gr-qc/9612019	gr-qc/9608035	gr-qc/9606039
gr-qc/9412053	gr-qc/9212005	gr-qc/0701100	gr-qc/0701087	gr-qc/0701040 ‡	gr-qc/0701012
gr-qc/0612163	gr-qc/0612159	gr-qc/0612053	gr-qc/0610104	gr-qc/0610083	gr-qc/0610029
gr-qc/0609038	gr-qc/0608054	gr-qc/0607125	gr-qc/0606058	gr-qc/0605046	gr-qc/0603134
gr-qc/9911007	gr-qc/0603128	gr-qc/0602095	gr-qc/0512120	gr-qc/0512109	gr-qc/0511082
gr-qc/0509093	gr-qc/0507104	gr-qc/0507090	gr-qc/0506108	gr-qc/0505035	gr-qc/0505031
gr-qc/0412096	gr-qc/0411062	gr-qc/0410026	gr-qc/0409059	gr-qc/0409023	gr-qc/0408026
gr-qc/0407083	gr-qc/0308054	gr-qc/0305086	gr-qc/0304017	gr-qc/0303047	gr-qc/0303031
gr-qc/0302108 ‡	gr-qc/0212037	gr-qc/0211015	gr-qc/0210079	gr-qc/0206043	gr-qc/0205106
gr-qc/0205087	gr-qc/0112065	gr-qc/0112044	gr-qc/0111107	gr-qc/0111070	gr-qc/0110102
gr-qc/0106050	gr-qc/0106049	gr-qc/0103013	gr-qc/0103009	gr-qc/0009008	gr-qc/0008005
gr-qc/0006051	gr-qc/0006048	astro-ph/9912558	astro-ph/9912554	astro-ph/9912549	astro-ph/9912548
astro-ph/9912424	astro-ph/9912343	astro-ph/9912211	astro-ph/9912166	astro-ph/9912140	astro-ph/9912139
astro-ph/9911518	astro-ph/9911372	astro-ph/9911260	astro-ph/9911246	astro-ph/9910566	astro-ph/9910545
astro-ph/9910459	astro-ph/9910396	astro-ph/9910359	astro-ph/9910350	astro-ph/9910315	astro-ph/9910266
astro-ph/9910265	astro-ph/9910207	astro-ph/9910187	astro-ph/9910182	astro-ph/9910166	astro-ph/9910097
astro-ph/9910031	astro-ph/9910004	astro-ph/9909478	astro-ph/9909437	astro-ph/9909389	astro-ph/9909386
astro-ph/9909321	astro-ph/9909280	astro-ph/9909279	astro-ph/9909252	astro-ph/9909226	astro-ph/9909124
astro-ph/9909087	astro-ph/9909068	astro-ph/9909064	astro-ph/9909012	astro-ph/9908335	astro-ph/9908332
astro-ph/9908305	astro-ph/9908213	astro-ph/9908152	astro-ph/9908114	astro-ph/9908047	astro-ph/9907409
astro-ph/9907337	astro-ph/9907326	astro-ph/9907292	astro-ph/9907260	astro-ph/9907165	astro-ph/9907140
astro-ph/9906481	astro-ph/9906391	astro-ph/9906379	astro-ph/9906375	astro-ph/9906286	astro-ph/9906277
astro-ph/9906261	astro-ph/9906260	astro-ph/9906224	astro-ph/9906204	astro-ph/9906160	astro-ph/9906159
astro-ph/9906093	astro-ph/9906049	astro-ph/9906034	astro-ph/9906033	astro-ph/9905281	astro-ph/9905280
astro-ph/9905135	astro-ph/9905024	astro-ph/9904401	astro-ph/9904396	astro-ph/9904366	astro-ph/9904365
astro-ph/9904317	astro-ph/9904291	astro-ph/9904284	astro-ph/9904283	astro-ph/9904263	astro-ph/9904251
astro-ph/9904159	astro-ph/9904064	astro-ph/9904001	astro-ph/9903465	astro-ph/9903420	astro-ph/9903182
astro-ph/9903003	astro-ph/9903002	astro-ph/9902210	astro-ph/9902172	astro-ph/9902014	astro-ph/9901358

Continued on next page

astro-ph/9901340	astro-ph/9901313	astro-ph/9901242	astro-ph/9901213	astro-ph/9901185	astro-ph/9901178
astro-ph/9901145	astro-ph/9901143	astro-ph/9901138	astro-ph/9901122	astro-ph/9901109	astro-ph/9901058
astro-ph/9812328	astro-ph/9812290	astro-ph/9812277	astro-ph/9812242	astro-ph/9812241	astro-ph/9812211
astro-ph/9812117	astro-ph/9812026	astro-ph/9812022	astro-ph/9812015	astro-ph/9812013	astro-ph/9811477
astro-ph/9811454	astro-ph/9811434	astro-ph/9811324	astro-ph/9811312	astro-ph/9811290	astro-ph/9811143
astro-ph/9811095	astro-ph/9811019	astro-ph/9811010	astro-ph/9810403	astro-ph/9810389	astro-ph/9810341
astro-ph/9810277	astro-ph/9810204	astro-ph/9810130	astro-ph/9810092	astro-ph/9810076	astro-ph/9809397
astro-ph/9809366	astro-ph/9809023	astro-ph/9808220	astro-ph/9808204	astro-ph/9808138	astro-ph/9808072
astro-ph/9808052	astro-ph/9808032	astro-ph/9807347	astro-ph/9807236	astro-ph/9807177	astro-ph/9807146
astro-ph/9807122	astro-ph/9807091	astro-ph/9807034	astro-ph/9806362	astro-ph/9806304	astro-ph/9806289
astro-ph/9806261	astro-ph/9806257	astro-ph/9806198	astro-ph/9806196	astro-ph/9806195	astro-ph/9806165
astro-ph/9806072	astro-ph/9806071	astro-ph/9805346	astro-ph/9805319	astro-ph/9805317	astro-ph/9805277
astro-ph/9805273	astro-ph/9805142	astro-ph/9805120	astro-ph/9804295	astro-ph/9804255	astro-ph/9804250
astro-ph/9804057	astro-ph/9804053	astro-ph/9804050	astro-ph/9803328	astro-ph/9803281	astro-ph/9803082
astro-ph/9803061	astro-ph/9802215	astro-ph/9802142	astro-ph/9802111	astro-ph/9802005	astro-ph/9801314
astro-ph/9801290	astro-ph/9801234	astro-ph/9801192	astro-ph/9801131	astro-ph/9801123	astro-ph/9801116
astro-ph/9801107	astro-ph/9801073	astro-ph/9801072	astro-ph/9801047	astro-ph/9712323	astro-ph/9712318
astro-ph/9712222	astro-ph/9712179	astro-ph/9712114	astro-ph/9712080	astro-ph/9711350	astro-ph/9711288
astro-ph/9711259	astro-ph/9711180	astro-ph/9711139	astro-ph/9711105	astro-ph/9711081	astro-ph/9711039
astro-ph/9710335	astro-ph/9710252	astro-ph/9710125	astro-ph/9710090	astro-ph/9710078	astro-ph/9710061
astro-ph/9710039	astro-ph/9709221	astro-ph/9709220	astro-ph/9709051	astro-ph/9709010	astro-ph/9708235
astro-ph/9708222	astro-ph/9708191	astro-ph/9708176	astro-ph/9708136	astro-ph/9708067	astro-ph/9708009
astro-ph/9707294	astro-ph/9707285	astro-ph/9707240	astro-ph/9707212	astro-ph/9706262	astro-ph/9706087
astro-ph/9706085	astro-ph/9705210	astro-ph/9705145	astro-ph/9705038	astro-ph/9705029	astro-ph/9704226
astro-ph/9704115	astro-ph/9704088	astro-ph/9704033	astro-ph/9703112	astro-ph/9703078	astro-ph/9703057
astro-ph/9703027	astro-ph/9702236	astro-ph/9702165	astro-ph/9702082	astro-ph/9702081	astro-ph/9702038
astro-ph/9701239	astro-ph/9701215	astro-ph/9701208	astro-ph/9612214	astro-ph/9612156	astro-ph/9612122
astro-ph/9612099	astro-ph/9612018	astro-ph/9611232	astro-ph/9611185	astro-ph/9611137	astro-ph/9611080
astro-ph/9611078	astro-ph/9611065	astro-ph/9611059	astro-ph/9611053	astro-ph/9611040	astro-ph/9610263
astro-ph/9610249	astro-ph/9610216	astro-ph/9610215	astro-ph/9610206	astro-ph/9610192	astro-ph/9610078
astro-ph/9610070	astro-ph/9610059	astro-ph/9610053	astro-ph/9610031	astro-ph/9610010	astro-ph/9610005
astro-ph/9610003	astro-ph/9609194	astro-ph/9609115	astro-ph/9609081	astro-ph/9609068	astro-ph/9609040
astro-ph/9609022	astro-ph/9609016	astro-ph/9608069	astro-ph/9608051	astro-ph/9608043	astro-ph/9608035
astro-ph/9607150	astro-ph/9607143	astro-ph/9607142	astro-ph/9607062	astro-ph/9607061	astro-ph/9607055
astro-ph/9607040	astro-ph/9607037	astro-ph/9607022	astro-ph/9606151	astro-ph/9606134	astro-ph/9606132
astro-ph/9606100	astro-ph/9606094	astro-ph/9606024	astro-ph/9606012	astro-ph/9605198	astro-ph/9605174
astro-ph/9605111	astro-ph/9605068	astro-ph/9605057	astro-ph/9605001	astro-ph/9604176	astro-ph/9604138
astro-ph/9603156	astro-ph/9603150	astro-ph/9603132	astro-ph/9603081	astro-ph/9603074	astro-ph/9603035
astro-ph/9603026	astro-ph/9602145	astro-ph/9602105	astro-ph/9601188	astro-ph/9601145	astro-ph/9601134
astro-ph/9601122	astro-ph/9512130	astro-ph/9512127	astro-ph/9512102	astro-ph/9512054	astro-ph/9511147
astro-ph/9511115	astro-ph/9511087	astro-ph/9511066	astro-ph/9511055	astro-ph/9511041	astro-ph/9511036
astro-ph/9511022	astro-ph/9510147	astro-ph/9510104	astro-ph/9510099	astro-ph/9510098	astro-ph/9510089
astro-ph/9510023	astro-ph/9510016	astro-ph/9509149	astro-ph/9509147	astro-ph/9509106	astro-ph/9509105
astro-ph/9509075	astro-ph/9509064	astro-ph/9509010	astro-ph/9508107	astro-ph/9508025	astro-ph/9508020
astro-ph/9508013	astro-ph/9508009	astro-ph/9507074	astro-ph/9507051	astro-ph/9507008	astro-ph/9507006
astro-ph/9506110	astro-ph/9506091	astro-ph/9506088	astro-ph/9506059	astro-ph/9506051	astro-ph/9506041
astro-ph/9506016	astro-ph/9506004	astro-ph/9506003	astro-ph/9505143	astro-ph/9505124	astro-ph/9505076
astro-ph/9505058	astro-ph/9505055	astro-ph/9505050	astro-ph/9505031	astro-ph/9505029	astro-ph/9505014
astro-ph/9505002	astro-ph/9504082	astro-ph/9504081	astro-ph/9504061	astro-ph/9504052	astro-ph/9504041
astro-ph/9504014	astro-ph/9503088	astro-ph/9503056	astro-ph/9503051	astro-ph/9502100	astro-ph/9502062
astro-ph/9502060	astro-ph/9502052	astro-ph/9502033	astro-ph/9502019	astro-ph/9502018	astro-ph/9501091
astro-ph/9501066	astro-ph/9501046	astro-ph/9412088	astro-ph/9412054	astro-ph/9412049	astro-ph/9412011
astro-ph/9412009	astro-ph/9412007	astro-ph/9411082	astro-ph/9411079	astro-ph/9411073	astro-ph/9411038
astro-ph/9411020	astro-ph/9410059	astro-ph/9410022	astro-ph/9409091	astro-ph/9409074	astro-ph/9409034

Continued on next page

astro-ph/9409030	astro-ph/9408102	astro-ph/9408094	astro-ph/9408029	astro-ph/9408028	astro-ph/9408018
astro-ph/9408002	astro-ph/9407085	astro-ph/9407072	astro-ph/9407071	astro-ph/9407069	astro-ph/9407068
astro-ph/9407013	astro-ph/9407012	astro-ph/9407011	astro-ph/9407004	astro-ph/9405010	astro-ph/9405003
astro-ph/9404011	astro-ph/9403065	astro-ph/9403012	astro-ph/9402028	astro-ph/9402027	astro-ph/9402017
astro-ph/9312037	astro-ph/9312020	astro-ph/9312011	astro-ph/9312008	astro-ph/9311077	astro-ph/9311073
astro-ph/9311049	astro-ph/9311046	astro-ph/9311044	astro-ph/9311043	astro-ph/9311018	astro-ph/9310014
astro-ph/9309039	astro-ph/9309007	astro-ph/9308022	astro-ph/9308021	astro-ph/9308006	astro-ph/9306004
astro-ph/9305011	astro-ph/9305010	astro-ph/9304017	astro-ph/9304004	astro-ph/9303019	astro-ph/0703783
astro-ph/0703778	astro-ph/0703757	astro-ph/0703704	astro-ph/0703673	astro-ph/0703624	astro-ph/0703590
astro-ph/0703512	astro-ph/0703471	astro-ph/0703466	astro-ph/0703462	astro-ph/0703430	astro-ph/0703416
astro-ph/0703370	astro-ph/0703365	astro-ph/0703362	astro-ph/0703352	astro-ph/0703348	astro-ph/0703342
astro-ph/0703337	astro-ph/0703308	astro-ph/0703262	astro-ph/0703259	astro-ph/0703195	astro-ph/0703183
astro-ph/0703115	astro-ph/0703004	astro-ph/0702654	astro-ph/0702587	astro-ph/0702586	astro-ph/0702575
astro-ph/0702568	astro-ph/0702546	astro-ph/0702501	astro-ph/0702495	astro-ph/0702478	astro-ph/0702470
astro-ph/0702443	astro-ph/0702373	astro-ph/0702368	astro-ph/0702360	astro-ph/0702333	astro-ph/0702328
astro-ph/0702275	astro-ph/0702260	astro-ph/0702241	astro-ph/0702173	astro-ph/0702167	astro-ph/0702164
astro-ph/0702010	astro-ph/0701884	astro-ph/0701848	astro-ph/0701826	astro-ph/0701792	astro-ph/0701780
astro-ph/0701673	astro-ph/0701672	astro-ph/0701624	astro-ph/0701598	astro-ph/0701594	astro-ph/0701542
astro-ph/0701446	astro-ph/0701426	astro-ph/0701418	astro-ph/0701365	astro-ph/0701358	astro-ph/0701331
astro-ph/0701292	astro-ph/0701286	astro-ph/0701088	astro-ph/0701086	astro-ph/0612786	astro-ph/0612750
astro-ph/0612733	astro-ph/0612732	astro-ph/0612616	astro-ph/0612565	astro-ph/0612473	astro-ph/0612467
astro-ph/0612462	astro-ph/0612410	astro-ph/0612387	astro-ph/0612327	astro-ph/0612253	astro-ph/0612239
astro-ph/0612219	astro-ph/0612130	astro-ph/0611948	astro-ph/0611925	astro-ph/0611921	astro-ph/0611913
astro-ph/0611887	astro-ph/0611872	astro-ph/0611864	astro-ph/0611783	astro-ph/0611684	astro-ph/0611664
astro-ph/0611582	astro-ph/0611506	astro-ph/0611496	astro-ph/0611494	astro-ph/0611370	astro-ph/0611353
astro-ph/0611303	astro-ph/0611221	astro-ph/0611205	astro-ph/0611168	astro-ph/0611144	astro-ph/0611129
astro-ph/0611124	astro-ph/0611113	astro-ph/0610961	astro-ph/0610922	astro-ph/0610821	astro-ph/0610806
astro-ph/0610731	astro-ph/0610727	astro-ph/0610682	astro-ph/0610674	astro-ph/0610618	astro-ph/0610428
astro-ph/0610425	astro-ph/0610336	astro-ph/0610280	astro-ph/0610269	astro-ph/0610249	astro-ph/0610134
astro-ph/0610056	astro-ph/0610038	astro-ph/0610034	astro-ph/0609782	astro-ph/0609777	astro-ph/0609714
astro-ph/0609713	astro-ph/0609687	astro-ph/0609652	astro-ph/0609640	astro-ph/0609629	astro-ph/0609572
astro-ph/0609510	astro-ph/0609500	astro-ph/0609425	astro-ph/0609413	astro-ph/0609388	astro-ph/0609375
astro-ph/0609361	astro-ph/0609326	astro-ph/0609126	astro-ph/0609125	astro-ph/0609072	astro-ph/0608706
astro-ph/0608690	astro-ph/0608661	astro-ph/0608637	astro-ph/0608634	astro-ph/0608614	astro-ph/0608613
astro-ph/0608607	astro-ph/0608602	astro-ph/0608562	astro-ph/0608535	astro-ph/0608528	astro-ph/0608526
astro-ph/0608523	astro-ph/0608407	astro-ph/0608390	astro-ph/0608385	astro-ph/0608304	astro-ph/0608276
astro-ph/0608228	astro-ph/0608175	astro-ph/0608151	astro-ph/0607639	astro-ph/0607621	astro-ph/0607555
astro-ph/0607437	astro-ph/0607411	astro-ph/0607394	astro-ph/0607391	astro-ph/0607374	astro-ph/0607341
astro-ph/0607327	astro-ph/0607319	astro-ph/0607142	astro-ph/0607131	astro-ph/0607126	astro-ph/0607121
astro-ph/0607100	astro-ph/0607073	astro-ph/0607042	astro-ph/0606699	astro-ph/0606654	astro-ph/0606566
astro-ph/0606483	astro-ph/0606482	astro-ph/0606447	astro-ph/0606435	astro-ph/0606429	astro-ph/0606415
astro-ph/0606360	astro-ph/0606350	astro-ph/0606315	astro-ph/0606281	astro-ph/0606208	astro-ph/0606197
astro-ph/0606190	astro-ph/0606073	astro-ph/0606028	astro-ph/0606015	astro-ph/0606010	astro-ph/0605724
astro-ph/0605719	astro-ph/0605706	astro-ph/0605697	astro-ph/0605688	astro-ph/0605637	astro-ph/0605630
astro-ph/0605500	astro-ph/0605424	astro-ph/0605322	astro-ph/0605271	astro-ph/0605254	astro-ph/0605249
astro-ph/0605205	astro-ph/0605102	astro-ph/0604587	astro-ph/0604518	astro-ph/0604506	astro-ph/0604418
astro-ph/0604393	astro-ph/0604311	astro-ph/0604307	astro-ph/0604126	astro-ph/0604024	astro-ph/0603796
astro-ph/0603775	astro-ph/0603704	astro-ph/0603661	astro-ph/0603660	astro-ph/0603602	astro-ph/0603541
astro-ph/0603387	astro-ph/0603368	astro-ph/0602632	astro-ph/0602584	astro-ph/0602440	astro-ph/0602400
astro-ph/0602394	astro-ph/0602349	astro-ph/0602325	astro-ph/0602266	astro-ph/0602161	astro-ph/0601669
astro-ph/0601602	astro-ph/0601581	astro-ph/0601489	astro-ph/0601431	astro-ph/0601422	astro-ph/0601404
astro-ph/0601344	astro-ph/0601301	astro-ph/0601298	astro-ph/0601274	astro-ph/0601266	astro-ph/0601249
astro-ph/0601248	astro-ph/0601233	astro-ph/0512631	astro-ph/0512569	astro-ph/0512509	astro-ph/0512507
astro-ph/0512494	astro-ph/0512482	astro-ph/0512454	astro-ph/0512425	astro-ph/0512405	astro-ph/0512386

Continued on next page

astro-ph/0512381	astro-ph/0512309	astro-ph/0512281	astro-ph/0512217	astro-ph/0512156	astro-ph/0512106
astro-ph/0512067	astro-ph/0512056	astro-ph/0511805	astro-ph/0511796	astro-ph/0511789	astro-ph/0511768
astro-ph/0511713	astro-ph/0511692	astro-ph/0511687	astro-ph/0511675	astro-ph/0511530	astro-ph/0511494
astro-ph/0511357	astro-ph/0511262	astro-ph/0511164	astro-ph/0511143	astro-ph/0511085	astro-ph/0511043
astro-ph/0510839	astro-ph/0510743	astro-ph/0510722	astro-ph/0510714	astro-ph/0510656	astro-ph/0510632
astro-ph/0510628	astro-ph/0510592	astro-ph/0510583	astro-ph/0510576	astro-ph/0510575	astro-ph/0510557
astro-ph/0510490	astro-ph/0510439	astro-ph/0510123	astro-ph/0510117	astro-ph/0510088	astro-ph/0510031
astro-ph/0510024	astro-ph/0509897	astro-ph/0509858	astro-ph/0509856	astro-ph/0509810	astro-ph/0509799
astro-ph/0509755	astro-ph/0509680	astro-ph/0509592	astro-ph/0509590	astro-ph/0509565	astro-ph/0509532
astro-ph/0509519	astro-ph/0509512	astro-ph/0509417	astro-ph/0509382	astro-ph/0509320	astro-ph/0509318
astro-ph/0509269	astro-ph/0509259	astro-ph/0509196	astro-ph/0509072	astro-ph/0508668	astro-ph/0508666
astro-ph/0508639	astro-ph/0508635	astro-ph/0508624	astro-ph/0508617	astro-ph/0508572	astro-ph/0508553
astro-ph/0508531	astro-ph/0508508	astro-ph/0508497	astro-ph/0508488	astro-ph/0508419	astro-ph/0508413
astro-ph/0508367	astro-ph/0508349	astro-ph/0508279	astro-ph/0508272	astro-ph/0508263	astro-ph/0508226
astro-ph/0508215	astro-ph/0508160	astro-ph/0508159	astro-ph/0508153	astro-ph/0508141	astro-ph/0508053
astro-ph/0508049	astro-ph/0508048	astro-ph/0507707	astro-ph/0507589	astro-ph/0507575	astro-ph/0507399
astro-ph/0507300	astro-ph/0507222	astro-ph/0507197	astro-ph/0507142	astro-ph/0507108	astro-ph/0506740
astro-ph/0506732	astro-ph/0506676	astro-ph/0506627	astro-ph/0506623	astro-ph/0506571	astro-ph/0506538
astro-ph/0506520	astro-ph/0506447	astro-ph/0506432	astro-ph/0506395	astro-ph/0506389	astro-ph/0506345
astro-ph/0506219	astro-ph/0506103	astro-ph/0506070	astro-ph/0506061	astro-ph/0506009	astro-ph/0505626
astro-ph/0505619	astro-ph/0505605	astro-ph/0505602	astro-ph/0505579	astro-ph/0505497	astro-ph/0505414
astro-ph/0505369	astro-ph/0505356	astro-ph/0505313	astro-ph/0505290	astro-ph/0505266	astro-ph/0505237
astro-ph/0505179	astro-ph/0505142	astro-ph/0505131	astro-ph/0505095	astro-ph/0505058	astro-ph/0505017
astro-ph/0504631	astro-ph/0504629	astro-ph/0504623	astro-ph/0504581	astro-ph/0504571	astro-ph/0504557
astro-ph/0504512	astro-ph/0504466	astro-ph/0504465	astro-ph/0504422	astro-ph/0504334	astro-ph/0504275
astro-ph/0504241	astro-ph/0504239	astro-ph/0504224	astro-ph/0504223	astro-ph/0504130	astro-ph/0504121
astro-ph/0504112	astro-ph/0504097	astro-ph/0504096	astro-ph/0504094	astro-ph/0504051	astro-ph/0503712
astro-ph/0503622	astro-ph/0503609	astro-ph/0503583	astro-ph/0503535	astro-ph/0503486	astro-ph/0503436
astro-ph/0503391	astro-ph/0503380	astro-ph/0503323	astro-ph/0503265	astro-ph/0503146	astro-ph/0503104
astro-ph/0503036	astro-ph/0503006	astro-ph/0502572	astro-ph/0502466	astro-ph/0502279	astro-ph/0502215
astro-ph/0502208	astro-ph/0502166	astro-ph/0502119	astro-ph/0502118	astro-ph/0502049	astro-ph/0502037
astro-ph/0501622	astro-ph/0501584	astro-ph/0501571	astro-ph/0501562	astro-ph/0501555	astro-ph/0501442
astro-ph/0501423	astro-ph/0501412	astro-ph/0501378	astro-ph/0501366	astro-ph/0501318	astro-ph/0501273
astro-ph/0501055	astro-ph/0501050	astro-ph/0412652	astro-ph/0412630	astro-ph/0412624	astro-ph/0412615
astro-ph/0412614	astro-ph/0412586	astro-ph/0412442	astro-ph/0412441	astro-ph/0412419	astro-ph/0412413
astro-ph/0412322	astro-ph/0412273	astro-ph/0412208	astro-ph/0412170	astro-ph/0412169	astro-ph/0412139
astro-ph/0412123	astro-ph/0412103	astro-ph/0412089	astro-ph/0412087	astro-ph/0412061	astro-ph/0412059
astro-ph/0412035	astro-ph/0412018	astro-ph/0411795	astro-ph/0411694	astro-ph/0411629	astro-ph/0411586
astro-ph/0411548	astro-ph/0411529	astro-ph/0411515	astro-ph/0411503	astro-ph/0411491	astro-ph/0411454
astro-ph/0411452	astro-ph/0411409	astro-ph/0411396	astro-ph/0411344	astro-ph/0411292	astro-ph/0411262
astro-ph/0411244	astro-ph/0411215	astro-ph/0411039	astro-ph/0410621	astro-ph/0410591	astro-ph/0410573
astro-ph/0410470	astro-ph/0410359	astro-ph/0410338	astro-ph/0410114	astro-ph/0410029	astro-ph/0410028
astro-ph/0409740	astro-ph/0409633	astro-ph/0409630	astro-ph/0409629	astro-ph/0409606	astro-ph/0409605
astro-ph/0409565	astro-ph/0409563	astro-ph/0409549	astro-ph/0409530	astro-ph/0409403	astro-ph/0409353
astro-ph/0409320	astro-ph/0409305	astro-ph/0409237	astro-ph/0409201	astro-ph/0409162	astro-ph/0409145
astro-ph/0409121	astro-ph/0409064	astro-ph/0409027	astro-ph/0408577	astro-ph/0408573	astro-ph/0408478
astro-ph/0408346	astro-ph/0408341	astro-ph/0408272	astro-ph/0408204	astro-ph/0408192	astro-ph/0408184
astro-ph/0408146	astro-ph/0408006	astro-ph/0407646	astro-ph/0407623	astro-ph/0407575	astro-ph/0407532
astro-ph/0407522	astro-ph/0407428	astro-ph/0407418	astro-ph/0407321	astro-ph/0407288	astro-ph/0407245
astro-ph/0407239	astro-ph/0407207	astro-ph/0407117	astro-ph/0407111	astro-ph/0406673	astro-ph/0406585
astro-ph/0406541	astro-ph/0406537	astro-ph/0406533	astro-ph/0406531	astro-ph/0406514	astro-ph/0406491
astro-ph/0406487	astro-ph/0406417	astro-ph/0406297	astro-ph/0406285	astro-ph/0406282	astro-ph/0406247
astro-ph/0406241	astro-ph/0406204	astro-ph/0406194	astro-ph/0406174	astro-ph/0406152	astro-ph/0406139
astro-ph/0406126	astro-ph/0406114	astro-ph/0406079	astro-ph/0406034	astro-ph/0405625	astro-ph/0405623

Continued on next page

astro-ph/0405598	astro-ph/0405572	astro-ph/0405508	astro-ph/0405496	astro-ph/0405491	astro-ph/0405479
astro-ph/0405466	astro-ph/0405456	astro-ph/0405455	astro-ph/0405442	astro-ph/0405371	astro-ph/0405363
astro-ph/0405342	astro-ph/0405266	astro-ph/0405242	astro-ph/0405235	astro-ph/0405231	astro-ph/0405216
astro-ph/0405189	astro-ph/0405033	astro-ph/0404600	astro-ph/0404568	astro-ph/0404499	astro-ph/0404490
astro-ph/0404483	astro-ph/0404465	astro-ph/0404311	astro-ph/0404205	astro-ph/0404117	astro-ph/0404086
astro-ph/0404033	astro-ph/0403698	astro-ph/0403694	astro-ph/0403619	astro-ph/0403614	astro-ph/0403596
astro-ph/0403571	astro-ph/0403528	astro-ph/0403514	astro-ph/0403417	astro-ph/0403384	astro-ph/0403372
astro-ph/0403352	astro-ph/0403324	astro-ph/0403322	astro-ph/0403294	astro-ph/0403229	astro-ph/0403206
astro-ph/0403164	astro-ph/0403154	astro-ph/0403102	astro-ph/0403077	astro-ph/0403064	astro-ph/0403048
astro-ph/0402634	astro-ph/0402588	astro-ph/0402525	astro-ph/0402516	astro-ph/0402504	astro-ph/0402479
astro-ph/0402405	astro-ph/0402390	astro-ph/0402366	astro-ph/0402346	astro-ph/0402316	astro-ph/0402210
astro-ph/0402157	astro-ph/0402095	astro-ph/0402055	astro-ph/0402045	astro-ph/0402033	astro-ph/0401609
astro-ph/0401591	astro-ph/0401575	astro-ph/0401565	astro-ph/0401512	astro-ph/0401511	astro-ph/0401378
astro-ph/0401341	astro-ph/0401185	astro-ph/0401162	astro-ph/0401140	astro-ph/0401097	astro-ph/0401088
astro-ph/0312651	astro-ph/0312645	astro-ph/0312606	astro-ph/0312605	astro-ph/0312570	astro-ph/0312547
astro-ph/0312544	astro-ph/0312358	astro-ph/0312243	astro-ph/0312221	astro-ph/0312194	astro-ph/0312109
astro-ph/0312086	astro-ph/0312072	astro-ph/0312002	astro-ph/0311631	astro-ph/0311614	astro-ph/0311594
astro-ph/0311522	astro-ph/0311361	astro-ph/0311348	astro-ph/0311259	astro-ph/0311240	astro-ph/0311185
astro-ph/0311150	astro-ph/0311145	astro-ph/0311100	astro-ph/0311083	astro-ph/0311052	astro-ph/0311049
astro-ph/0311020	astro-ph/0310908	astro-ph/0310897	astro-ph/0310798	astro-ph/0310756	astro-ph/0310707
astro-ph/0310703	astro-ph/0310666	astro-ph/0310638	astro-ph/0310579	astro-ph/0310572	astro-ph/0310534
astro-ph/0310531	astro-ph/0310528	astro-ph/0310439	astro-ph/0310376	astro-ph/0310334	astro-ph/0310294
astro-ph/0310219	astro-ph/0310193	astro-ph/0310192	astro-ph/0310005	astro-ph/0309812	astro-ph/0309762
astro-ph/0309755	astro-ph/0309738	astro-ph/0309735	astro-ph/0309686	astro-ph/0309652	astro-ph/0309617
astro-ph/0309611	astro-ph/0309517	astro-ph/0309490	astro-ph/0309465	astro-ph/0309412	astro-ph/0309405
astro-ph/0309343	astro-ph/0309330	astro-ph/0309329	astro-ph/0309303	astro-ph/0309273	astro-ph/0309163
astro-ph/0309020	astro-ph/0308518	astro-ph/0308515	astro-ph/0308472	astro-ph/0308385	astro-ph/0308348
astro-ph/0308183	astro-ph/0308092	astro-ph/0308054	astro-ph/0308007	astro-ph/0307524	astro-ph/0307465
astro-ph/0307437	astro-ph/0307403	astro-ph/0307358	astro-ph/0307350	astro-ph/0307316	astro-ph/0307270
astro-ph/0307214	astro-ph/0307209	astro-ph/0307206	astro-ph/0307190	astro-ph/0307184	astro-ph/0307154
astro-ph/0307141	astro-ph/0307115	astro-ph/0307082	astro-ph/0307042	astro-ph/0307026	astro-ph/0306561
astro-ph/0306520	astro-ph/0306515	astro-ph/0306493	astro-ph/0306490	astro-ph/0306445	astro-ph/0306437
astro-ph/0306413	astro-ph/0306402	astro-ph/0306385	astro-ph/0306374	astro-ph/0306365	astro-ph/0306327
astro-ph/0306286	astro-ph/0306282	astro-ph/0306233	astro-ph/0306205	astro-ph/0306203	astro-ph/0306134
astro-ph/0306124	astro-ph/0306102	astro-ph/0306081	astro-ph/0306020	astro-ph/0305584	astro-ph/0305547
astro-ph/0305516	astro-ph/0305496	astro-ph/0305310	astro-ph/0305300	astro-ph/0305187	astro-ph/0305075
astro-ph/0305025	astro-ph/0304504	astro-ph/0304474	astro-ph/0304418	astro-ph/0304385	astro-ph/0304375
astro-ph/0304355	astro-ph/0304317	astro-ph/0304315	astro-ph/0304246	astro-ph/0304183	astro-ph/0304029
astro-ph/0303622	astro-ph/0303610	astro-ph/0303564	astro-ph/0303524	astro-ph/0303455	astro-ph/0303440
astro-ph/0303422	astro-ph/0303349	astro-ph/0303313	astro-ph/0303262	astro-ph/0303238	astro-ph/0303112
astro-ph/0303058	astro-ph/0302461	astro-ph/0302445	astro-ph/0302444	astro-ph/0302443	astro-ph/0302335
astro-ph/0302318	astro-ph/0302257	astro-ph/0302071	astro-ph/0302035	astro-ph/0302031	astro-ph/0302030
astro-ph/0302005	astro-ph/0301640	astro-ph/0301612	astro-ph/0301551	astro-ph/0301533	astro-ph/0301505
astro-ph/0301503	astro-ph/0301490	astro-ph/0301399	astro-ph/0301365	astro-ph/0301360	astro-ph/0301341
astro-ph/0301301	astro-ph/0301271	astro-ph/0301270	astro-ph/0301242	astro-ph/0301229	astro-ph/0212575
astro-ph/0212560	astro-ph/0212555	astro-ph/0212465	astro-ph/0212438	astro-ph/0212390	astro-ph/0212293
astro-ph/0212290	astro-ph/0212275	astro-ph/0212268	astro-ph/0212266	astro-ph/0212251	astro-ph/0212176
astro-ph/0212146	astro-ph/0212131	astro-ph/0212114	astro-ph/0212102	astro-ph/0212049	astro-ph/0212009
astro-ph/0211573	astro-ph/0211559	astro-ph/0211535	astro-ph/0211500	astro-ph/0211354	astro-ph/0211327
astro-ph/0211323	astro-ph/0211303	astro-ph/0211302	astro-ph/0211294	astro-ph/0211239	astro-ph/0211238
astro-ph/0210658	astro-ph/0210617	astro-ph/0210599	astro-ph/0210495	astro-ph/0210441	astro-ph/0210184
astro-ph/0210079	astro-ph/0210052	astro-ph/0209304	astro-ph/0209291	astro-ph/0209205	astro-ph/0209128
astro-ph/0209121	astro-ph/0208458	astro-ph/0208419	astro-ph/0208403	astro-ph/0208268	astro-ph/0208257
astro-ph/0208093	astro-ph/0208090	astro-ph/0208045	astro-ph/0207673	astro-ph/0207670	astro-ph/0207628

Continued on next page

astro-ph/0207557	astro-ph/0207534	astro-ph/0207487	astro-ph/0207469	astro-ph/0207467	astro-ph/0207423
astro-ph/0207349	astro-ph/0207308	astro-ph/0207299	astro-ph/0207297	astro-ph/0207231	astro-ph/0207125
astro-ph/0207113	astro-ph/0207049	astro-ph/0207048	astro-ph/0207045	astro-ph/0206477	astro-ph/0206455
astro-ph/0206349	astro-ph/0206346	astro-ph/0206345	astro-ph/0206304	astro-ph/0206247	astro-ph/0206176
astro-ph/0206126	astro-ph/0206125	astro-ph/0206036	astro-ph/0206026	astro-ph/0205469	astro-ph/0205464
astro-ph/0205420	astro-ph/0205412	astro-ph/0205406	astro-ph/0205391	astro-ph/0205359	astro-ph/0205335
astro-ph/0205316	astro-ph/0205281	astro-ph/0205216	astro-ph/0205205	astro-ph/0205199	astro-ph/0205158
astro-ph/0205146	astro-ph/0205140	astro-ph/0205066	astro-ph/0205023	astro-ph/0205016	astro-ph/0204521
astro-ph/0204425	astro-ph/0204411	astro-ph/0204375	astro-ph/0204307	astro-ph/0204293	astro-ph/0204218
astro-ph/0204164	astro-ph/0204108	astro-ph/0204028	astro-ph/0203520	astro-ph/0203507	astro-ph/0203500
astro-ph/0203495	astro-ph/0203488	astro-ph/0203457	astro-ph/0203448	astro-ph/0203429	astro-ph/0203381
astro-ph/0203293	astro-ph/0203267	astro-ph/0203242	astro-ph/0203240	astro-ph/0203200	astro-ph/0203012
astro-ph/0203008	astro-ph/0203007	astro-ph/0203004	astro-ph/0202302	astro-ph/0202250	astro-ph/0202244
astro-ph/0202099	astro-ph/0201520	astro-ph/0201376	astro-ph/0201269	astro-ph/0201212	astro-ph/0201154
astro-ph/0201153	astro-ph/0201152	astro-ph/0201146	astro-ph/0112561	astro-ph/0112550	astro-ph/0112522
astro-ph/0112454	astro-ph/0112393	astro-ph/0112336	astro-ph/0112272	astro-ph/0112255	astro-ph/0112072
astro-ph/0112069	astro-ph/0112023	astro-ph/0111536	astro-ph/0111530	astro-ph/0111366	astro-ph/0111329
astro-ph/0111325	astro-ph/0111292	astro-ph/0111288	astro-ph/0111230	astro-ph/0111196	astro-ph/0111176
astro-ph/0111158	astro-ph/0111137	astro-ph/0111117	astro-ph/0111069	astro-ph/0111053	astro-ph/0111005
astro-ph/0110690	astro-ph/0110680	astro-ph/0110632	astro-ph/0110601	astro-ph/0110578	astro-ph/0110561
astro-ph/0110484	astro-ph/0110431	astro-ph/0110427	astro-ph/0110313	astro-ph/0110225	astro-ph/0110073
astro-ph/0110061	astro-ph/0109555	astro-ph/0109499	astro-ph/0109451	astro-ph/0109450	astro-ph/0109432
astro-ph/0109392	astro-ph/0109324	astro-ph/0109318	astro-ph/0108488	astro-ph/0108422	astro-ph/0108327
astro-ph/0108319	astro-ph/0108288	astro-ph/0108224	astro-ph/0108203	astro-ph/0108172	astro-ph/0108151
astro-ph/0108116	astro-ph/0108108	astro-ph/0108073	astro-ph/0108050	astro-ph/0108014	astro-ph/0107567
astro-ph/0107550	astro-ph/0107490	astro-ph/0107479	astro-ph/0107442	astro-ph/0107423	astro-ph/0107397
astro-ph/0107284	astro-ph/0107239	astro-ph/0107210	astro-ph/0107079	astro-ph/0107057	astro-ph/0106542
astro-ph/0106540	astro-ph/0106458	astro-ph/0106380	astro-ph/0106314	astro-ph/0106296	astro-ph/0106268
astro-ph/0106251	astro-ph/0106228	astro-ph/0106108	astro-ph/0106100	astro-ph/0106099	astro-ph/0106048
astro-ph/0106008	astro-ph/0106002	astro-ph/0105564	astro-ph/0105443	astro-ph/0105424	astro-ph/0105349
astro-ph/0105328	astro-ph/0105318	astro-ph/0105316	astro-ph/0105248	astro-ph/0105232	astro-ph/0105184
astro-ph/0105156	astro-ph/0105152	astro-ph/0105048	astro-ph/0104465	astro-ph/0104293	astro-ph/0104255
astro-ph/0104221	astro-ph/0104200	astro-ph/0104192	astro-ph/0104173	astro-ph/0104110	astro-ph/0104106
astro-ph/0104026	astro-ph/0104013	astro-ph/0104002	astro-ph/0103452	astro-ph/0103122	astro-ph/0103050
astro-ph/0102466	astro-ph/0102419	astro-ph/0102400	astro-ph/0102389	astro-ph/0102349	astro-ph/0102334
astro-ph/0102304	astro-ph/0102082	astro-ph/0101528	astro-ph/0101524	astro-ph/0101480	astro-ph/0101413
astro-ph/0101349	astro-ph/0101227	astro-ph/0101204	astro-ph/0101134	astro-ph/0012504	astro-ph/0012346
astro-ph/0012337	astro-ph/0012234	astro-ph/0012233	astro-ph/0012178	astro-ph/0012161	astro-ph/0012018
astro-ph/0011506	astro-ph/0011376	astro-ph/0011333	astro-ph/0011295	astro-ph/0011292	astro-ph/0011079
astro-ph/0011042	astro-ph/0011017	astro-ph/0010616	astro-ph/0010598	astro-ph/0010569	astro-ph/0010536
astro-ph/0010525	astro-ph/0010436	astro-ph/0010389	astro-ph/0010358	astro-ph/0010226	astro-ph/0010223
astro-ph/0010135	astro-ph/0010056	astro-ph/0010050	astro-ph/0009407	astro-ph/0009317	astro-ph/0009306
astro-ph/0009254	astro-ph/0009197	astro-ph/0009167	astro-ph/0009125	astro-ph/0009074	astro-ph/0009003
astro-ph/0008451	astro-ph/0008422	astro-ph/0008339	astro-ph/0008234	astro-ph/0008157	astro-ph/0008156
astro-ph/0008115	astro-ph/0008104	astro-ph/0008046	astro-ph/0008028	astro-ph/0007444	astro-ph/0007165
astro-ph/0007121	astro-ph/0007120	astro-ph/0007067	astro-ph/0007045	astro-ph/0006453	astro-ph/0006393
astro-ph/0006344	astro-ph/0006343	astro-ph/0006316	astro-ph/0006310	astro-ph/0006281	astro-ph/0006275
astro-ph/0006203	astro-ph/0006134	astro-ph/0006074	astro-ph/0006048	astro-ph/0005568	astro-ph/0005504
astro-ph/0005381	astro-ph/0005340	astro-ph/0005332	astro-ph/0005299	astro-ph/0005265	astro-ph/0005260
astro-ph/0005210	astro-ph/0005206	astro-ph/0005194	astro-ph/0005095	astro-ph/0005001	astro-ph/0004397
astro-ph/0004381	astro-ph/0004373	astro-ph/0004356	astro-ph/0004352	astro-ph/0004342	astro-ph/0004332
astro-ph/0004292	astro-ph/0004150	astro-ph/0004115	astro-ph/0004046	astro-ph/0004037	astro-ph/0004023
astro-ph/0003483	astro-ph/0003434	astro-ph/0003388	astro-ph/0003365	astro-ph/0003350	astro-ph/0003302
astro-ph/0003205	astro-ph/0003199	astro-ph/0003105	astro-ph/0003046	astro-ph/0003018	astro-ph/0002495

Continued on next page

astro-ph/0002471	astro-ph/0002462	astro-ph/0002409	astro-ph/0002391	astro-ph/0002376	astro-ph/0002363
astro-ph/0002330	astro-ph/0002308	astro-ph/0002163	astro-ph/0002091	astro-ph/0002058	astro-ph/0002050
astro-ph/0001146	arxiv/0811.1942	arxiv/0811.1851	arxiv/0811.1582	arxiv/0811.1555	arxiv/0811.1270
arxiv/0811.1218	arxiv/0811.0967	arxiv/0811.0821	arxiv/0811.0698	arxiv/0811.0658	arxiv/0811.0477
arxiv/0811.0326	arxiv/0811.0250	arxiv/0811.0099	arxiv/0810.5560	arxiv/0810.5557	arxiv/0810.5483
arxiv/0810.5411	arxiv/0810.5401	arxiv/0810.5397	arxiv/0810.5304	arxiv/0810.5300	arxiv/0810.5292
arxiv/0810.5287	arxiv/0810.5167	arxiv/0810.5126	arxiv/0810.4925	arxiv/0810.4490	arxiv/0810.4446
arxiv/0810.4421	arxiv/0810.4311	arxiv/0810.4147	arxiv/0810.4110	arxiv/0810.4034	arxiv/0810.3676
arxiv/0810.3655	arxiv/0810.3614	arxiv/0810.3504	arxiv/0810.3435	arxiv/0810.3140	arxiv/0810.2827
arxiv/0810.2769	arxiv/0810.2301	arxiv/0810.2177	arxiv/0810.2119	arxiv/0810.2036	arxiv/0810.2022
arxiv/0810.1892	arxiv/0810.1724	arxiv/0810.1522	arxiv/0810.1502	arxiv/0810.1493	arxiv/0810.1472
arxiv/0810.1291	arxiv/0810.1245	arxiv/0810.0709	arxiv/0810.0703	arxiv/0810.0423	arxiv/0810.0291
arxiv/0810.0277	arxiv/0810.0274	arxiv/0809.5072	arxiv/0809.5057	arxiv/0809.4667	arxiv/0809.4506
arxiv/0809.4438	arxiv/0809.4186	arxiv/0809.4125	arxiv/0809.4065	arxiv/0809.4011	arxiv/0809.3906
arxiv/0809.3894	arxiv/0809.3465	arxiv/0809.3252	arxiv/0809.3196	arxiv/0809.3134	arxiv/0809.3113
arxiv/0809.2990	arxiv/0809.2942	arxiv/0809.2839	arxiv/0809.2783	arxiv/0809.2610	arxiv/0809.2601
arxiv/0809.2417	arxiv/0809.2269	arxiv/0809.2036	arxiv/0809.1976	arxiv/0809.1972	arxiv/0809.1871
arxiv/0809.1829	arxiv/0809.1653	arxiv/0809.1523	arxiv/0809.1519	arxiv/0809.1503	arxiv/0809.1322
arxiv/0809.1206	arxiv/0809.1159	arxiv/0809.0935	arxiv/0809.0901	arxiv/0809.0894	arxiv/0809.0677
arxiv/0809.0668	arxiv/0809.0497	arxiv/0809.0436	arxiv/0809.0375	arxiv/0809.0162	arxiv/0808.4151
arxiv/0808.4068 †	arxiv/0808.3968	arxiv/0808.3910	arxiv/0808.3902	arxiv/0808.3898	arxiv/0808.3867
arxiv/0808.3725	arxiv/0808.3607	arxiv/0808.3449	arxiv/0808.3425	arxiv/0808.3401	arxiv/0808.3384
arxiv/0808.2981	arxiv/0808.2823	arxiv/0808.2641	arxiv/0808.2595	arxiv/0808.2471	arxiv/0808.2318
arxiv/0808.2049	arxiv/0808.1727	arxiv/0808.1097	arxiv/0808.1050	arxiv/0808.0899	arxiv/0808.0881
arxiv/0808.0472	arxiv/0808.0332	arxiv/0808.0320	arxiv/0808.0196	arxiv/0808.0195	arxiv/0807.4642
arxiv/0807.4590	arxiv/0807.4529	arxiv/0807.4373	arxiv/0807.4363	arxiv/0807.4178	arxiv/0807.4175
arxiv/0807.3769	arxiv/0807.3736	arxiv/0807.3651	arxiv/0807.3429	arxiv/0807.3379	arxiv/0807.3343
arxiv/0807.3279	arxiv/0807.3029	arxiv/0807.3027	arxiv/0807.2926	arxiv/0807.2863	arxiv/0807.2753
arxiv/0807.2548	arxiv/0807.2250	arxiv/0807.2244	arxiv/0807.2102	arxiv/0807.1973	arxiv/0807.1508
arxiv/0807.1328	arxiv/0807.1218	arxiv/0807.1124	arxiv/0807.1020	arxiv/0807.0685	arxiv/0807.0646
arxiv/0807.0622	arxiv/0807.0232	arxiv/0806.4649	arxiv/0806.4099	arxiv/0806.3989	arxiv/0806.3767
arxiv/0806.3746	arxiv/0806.3689	arxiv/0806.3613	arxiv/0806.3610	arxiv/0806.3581	arxiv/0806.3434
arxiv/0806.3266	arxiv/0806.3225	arxiv/0806.3147	arxiv/0806.3108	arxiv/0806.3093	arxiv/0806.2986
arxiv/0806.2981	arxiv/0806.2911	arxiv/0806.2723	arxiv/0806.2681	arxiv/0806.2673	arxiv/0806.2662
arxiv/0806.2585	arxiv/0806.2320	arxiv/0806.2149	arxiv/0806.2065	arxiv/0806.1969	arxiv/0806.1962
arxiv/0806.1951	arxiv/0806.1923	arxiv/0806.1766	arxiv/0806.1707	arxiv/0806.1513	arxiv/0806.1501
arxiv/0806.1493	arxiv/0806.1191	arxiv/0806.0730	arxiv/0806.0683	arxiv/0806.0370	arxiv/0806.0232
arxiv/0806.0208	arxiv/0806.0116	arxiv/0806.0011	arxiv/0805.4818	arxiv/0805.4416	arxiv/0805.4400
arxiv/0805.4210	arxiv/0805.4163	arxiv/0805.4016	arxiv/0805.3945	arxiv/0805.3827	arxiv/0805.3540
arxiv/0805.3531	arxiv/0805.3423	arxiv/0805.3389	arxiv/0805.2895	arxiv/0805.2877	arxiv/0805.2836
arxiv/0805.2552	arxiv/0805.2431	arxiv/0805.2372	arxiv/0805.2270	arxiv/0805.2241	arxiv/0805.1926
arxiv/0805.1905	arxiv/0805.1823	arxiv/0805.1519	arxiv/0805.1244	arxiv/0805.1233	arxiv/0805.1133
arxiv/0805.1060	arxiv/0805.1054	arxiv/0805.0731	arxiv/0805.0494	arxiv/0805.0309	arxiv/0805.0124
arxiv/0804.4827	arxiv/0804.4596	arxiv/0804.4543	arxiv/0804.4518	arxiv/0804.4185	arxiv/0804.4172
arxiv/0804.3804	arxiv/0804.3792	arxiv/0804.3777	arxiv/0804.3601	arxiv/0804.3518	arxiv/0804.3350
arxiv/0804.3235	arxiv/0804.3141	arxiv/0804.2680	arxiv/0804.2544	arxiv/0804.2486	arxiv/0804.2207
arxiv/0804.2205	arxiv/0804.1976	arxiv/0804.1919	arxiv/0804.1854	arxiv/0804.1613	arxiv/0804.1588
arxiv/0804.1499	arxiv/0804.1191 †	arxiv/0804.0869	arxiv/0804.0863	arxiv/0804.0589	arxiv/0804.0426
arxiv/0804.0336	arxiv/0804.0294	arxiv/0804.0285	arxiv/0804.0282	arxiv/0804.0256	arxiv/0804.0232
arxiv/0804.0110	arxiv/0804.0024	arxiv/0804.0004	arxiv/0803.4477	arxiv/0803.4302	arxiv/0803.4196
arxiv/0803.3624	arxiv/0803.3223	arxiv/0803.3036	arxiv/0803.2706	arxiv/0803.2640	arxiv/0803.2164
arxiv/0803.2154	arxiv/0803.1954	arxiv/0803.1865	arxiv/0803.1585	arxiv/0803.1444	arxiv/0803.1431
arxiv/0803.1354	arxiv/0803.0977	arxiv/0803.0631	arxiv/0803.0468	arxiv/0803.0370	arxiv/0803.0157
arxiv/0803.0144	arxiv/0803.0002	arxiv/0802.4348	arxiv/0802.4336	arxiv/0802.3997	arxiv/0802.3847

Continued on next page

arxiv/0802.3830	arxiv/0802.3713	arxiv/0802.3530	arxiv/0802.3526	arxiv/0802.3384	arxiv/0802.3378
arxiv/0802.3245	arxiv/0802.3155	arxiv/0802.3127	arxiv/0802.2968	arxiv/0802.2922	arxiv/0802.2872
arxiv/0802.2534	arxiv/0802.2296	arxiv/0802.2265	arxiv/0802.2164	arxiv/0802.2123	arxiv/0802.2041
arxiv/0802.1934	arxiv/0802.1917	arxiv/0802.1775	arxiv/0802.1768	arxiv/0802.1726	arxiv/0802.1724
arxiv/0802.1628	arxiv/0802.1599	arxiv/0802.1163	arxiv/0802.1144	arxiv/0802.0941	arxiv/0802.0702
arxiv/0802.0599	arxiv/0802.0546	arxiv/0802.0506	arxiv/0802.0429	arxiv/0802.0392	arxiv/0802.0234
arxiv/0802.0052	arxiv/0801.4927	arxiv/0801.4895	arxiv/0801.4888	arxiv/0801.4728	arxiv/0801.4673
arxiv/0801.4517	arxiv/0801.4378	arxiv/0801.4284	arxiv/0801.4233	arxiv/0801.4135	arxiv/0801.4108
arxiv/0801.4103	arxiv/0801.4015	arxiv/0801.3847	arxiv/0801.3686	arxiv/0801.3459	arxiv/0801.3440
arxiv/0801.3271	arxiv/0801.3269	arxiv/0801.3241	arxiv/0801.3133	arxiv/0801.3029	arxiv/0801.2539
arxiv/0801.2527	arxiv/0801.2476	arxiv/0801.2169	arxiv/0801.2024	arxiv/0801.1984	arxiv/0801.1851
arxiv/0801.1831	arxiv/0801.1708	arxiv/0801.1662	arxiv/0801.1645	arxiv/0801.1624	arxiv/0801.1565
arxiv/0801.1442	arxiv/0801.1232	arxiv/0801.1156	arxiv/0801.1091	arxiv/0801.0990	arxiv/0801.0853
arxiv/0801.0740	arxiv/0801.0491	arxiv/0801.0201	arxiv/0801.0169	arxiv/0801.0167	arxiv/0712.4232
arxiv/0712.4206	arxiv/0712.4203	arxiv/0712.4190	arxiv/0712.4146	arxiv/0712.4100	arxiv/0712.4038
arxiv/0712.3800	arxiv/0712.3570	arxiv/0712.3499	arxiv/0712.3465	arxiv/0712.3151	arxiv/0712.2933
arxiv/0712.2675	arxiv/0712.2667	arxiv/0712.2576	arxiv/0712.2312	arxiv/0712.2280	arxiv/0712.1937
arxiv/0712.1816	arxiv/0712.1688	arxiv/0712.1645	arxiv/0712.1563	arxiv/0712.1476	arxiv/0712.1459
arxiv/0712.1234	arxiv/0712.0984	arxiv/0712.0816	arxiv/0712.0677	arxiv/0712.0562	arxiv/0712.0505
arxiv/0712.0486	arxiv/0712.0468	arxiv/0712.0253	arxiv/0712.0053	arxiv/0712.0019	arxiv/0712.0016
arxiv/0712.0007	arxiv/0711.5027	arxiv/0711.5022	arxiv/0711.4996	arxiv/0711.4989	arxiv/0711.4895
arxiv/0711.4866	arxiv/0711.4850	arxiv/0711.4621	arxiv/0711.4591	arxiv/0711.4587	arxiv/0711.4474
arxiv/0711.4077	arxiv/0711.3950	arxiv/0711.3881	arxiv/0711.3791	arxiv/0711.3788	arxiv/0711.3749
arxiv/0711.3682	arxiv/0711.3528	arxiv/0711.3341	arxiv/0711.3254	arxiv/0711.3148	arxiv/0711.3033
arxiv/0711.2906	arxiv/0711.2870	arxiv/0711.2858	arxiv/0711.2823	arxiv/0711.2791	arxiv/0711.2741
arxiv/0711.2698	arxiv/0711.2316	arxiv/0711.2302	arxiv/0711.1775	arxiv/0711.1707	arxiv/0711.1570
arxiv/0711.1506	arxiv/0711.1352	arxiv/0711.1297	arxiv/0711.1264	arxiv/0711.1240	arxiv/0711.1198
arxiv/0711.1105	arxiv/0711.0958	arxiv/0711.0488	arxiv/0711.0470	arxiv/0711.0466	arxiv/0710.5849
arxiv/0710.5833	arxiv/0710.5712	arxiv/0710.5603	arxiv/0710.5520	arxiv/0710.5517	arxiv/0710.5473
arxiv/0710.5431	arxiv/0710.5420	arxiv/0710.5311	arxiv/0710.5180	arxiv/0710.5121	arxiv/0710.5119
arxiv/0710.4966	arxiv/0710.4936	arxiv/0710.4922	arxiv/0710.4567	arxiv/0710.4384	arxiv/0710.4296
arxiv/0710.4136	arxiv/0710.4070	arxiv/0710.4029	arxiv/0710.3898	arxiv/0710.3685	arxiv/0710.3545
arxiv/0710.3169	arxiv/0710.3160	arxiv/0710.3018	arxiv/0710.2968	arxiv/0710.2618	arxiv/0710.2493
arxiv/0710.2458	arxiv/0710.2416	arxiv/0710.2360	arxiv/0710.2336	arxiv/0710.2287	arxiv/0710.2213
arxiv/0710.2104	arxiv/0710.2062	arxiv/0710.2058	arxiv/0710.2005	arxiv/0710.1856	arxiv/0710.1843
arxiv/0710.1759	arxiv/0710.1668	arxiv/0710.1561	arxiv/0710.1411	arxiv/0710.1198	arxiv/0710.1179
arxiv/0710.1108	arxiv/0710.1069	arxiv/0710.1044	arxiv/0710.1032	arxiv/0710.0449	arxiv/0710.0364
arxiv/0710.0224	arxiv/0710.0210	arxiv/0710.0135	arxiv/0709.4593	arxiv/0709.4576	arxiv/0709.4477
arxiv/0709.4027	arxiv/0709.3952	arxiv/0709.3918	arxiv/0709.3815	arxiv/0709.3376	arxiv/0709.3321
arxiv/0709.3301	arxiv/0709.3189	arxiv/0709.3114	arxiv/0709.2909	arxiv/0709.2561	arxiv/0709.2505
arxiv/0709.2486	arxiv/0709.2369	arxiv/0709.2301	arxiv/0709.2299	arxiv/0709.2297	arxiv/0709.2098
arxiv/0709.2078	arxiv/0709.1966	arxiv/0709.1862	arxiv/0709.1636	arxiv/0709.1632	arxiv/0709.1571
arxiv/0709.1510	arxiv/0709.1485	arxiv/0709.1218	arxiv/0709.1128	arxiv/0709.1106	arxiv/0709.0858
arxiv/0709.0745	arxiv/0709.0691	arxiv/0709.0462	arxiv/0709.0434	arxiv/0709.0327	arxiv/0709.0297
arxiv/0709.0174	arxiv/0709.0166	arxiv/0709.0131	arxiv/0709.0108	arxiv/0709.0046	arxiv/0709.0043
arxiv/0708.4363	arxiv/0708.4003	arxiv/0708.3983	arxiv/0708.3970	arxiv/0708.3790	arxiv/0708.3600
arxiv/0708.3371	arxiv/0708.3342	arxiv/0708.3179	arxiv/0708.2797	arxiv/0708.2768	arxiv/0708.2621
arxiv/0708.2618	arxiv/0708.2579	arxiv/0708.2370	arxiv/0708.2348	arxiv/0708.2166	arxiv/0708.2151
arxiv/0708.2030	arxiv/0708.1949	arxiv/0708.1891	arxiv/0708.1784	arxiv/0708.1492	arxiv/0708.1382
arxiv/0708.1376	arxiv/0708.1206	arxiv/0708.0762	arxiv/0708.0753	arxiv/0708.0247	arxiv/0707.4423
arxiv/0707.4375	arxiv/0707.4374	arxiv/0707.4247	arxiv/0707.4126	arxiv/0707.3813	arxiv/0707.3524
arxiv/0707.3334	arxiv/0707.3260	arxiv/0707.3049	arxiv/0707.2968	arxiv/0707.2960	arxiv/0707.2959
arxiv/0707.2470	arxiv/0707.2463	arxiv/0707.2377	arxiv/0707.2089	arxiv/0707.1949	arxiv/0707.1937
arxiv/0707.1758	arxiv/0707.1698	arxiv/0707.1536	arxiv/0707.1495	arxiv/0707.1488	arxiv/0707.1479

Continued on next page

arxiv/0707.1081	arxiv/0707.0737	arxiv/0707.0633	arxiv/0707.0628	arxiv/0707.0622	arxiv/0707.0618
arxiv/0707.0488	arxiv/0707.0472	arxiv/0707.0209	arxiv/0707.0196	arxiv/0706.4349	arxiv/0706.4198
arxiv/0706.4084	arxiv/0706.4071	arxiv/0706.3909	arxiv/0706.3895	arxiv/0706.3773	arxiv/0706.3703
arxiv/0706.3409	arxiv/0706.3357	arxiv/0706.3196	arxiv/0706.3149	arxiv/0706.3050	arxiv/0706.3019
arxiv/0706.2986	arxiv/0706.2775	arxiv/0706.2694	arxiv/0706.2558	arxiv/0706.2401	arxiv/0706.2237
arxiv/0706.2101	arxiv/0706.1976	arxiv/0706.1586	arxiv/0706.1357	arxiv/0706.1279	arxiv/0706.0974
arxiv/0706.0948	arxiv/0706.0918	arxiv/0706.0896	arxiv/0706.0875	arxiv/0706.0856	arxiv/0706.0852
arxiv/0706.0526	arxiv/0706.0208	arxiv/0706.0186	arxiv/0706.0039	arxiv/0706.0031	arxiv/0706.0006
arxiv/0705.4680	arxiv/0705.4633	arxiv/0705.4493	arxiv/0705.4477	arxiv/0705.4298	arxiv/0705.4219
arxiv/0705.4158	arxiv/0705.4056	arxiv/0705.4043	arxiv/0705.4027	arxiv/0705.3843	arxiv/0705.3655
arxiv/0705.3345	arxiv/0705.3017	arxiv/0705.2924	arxiv/0705.2908	arxiv/0705.2881	arxiv/0705.2846
arxiv/0705.2760	arxiv/0705.2695	arxiv/0705.2610	arxiv/0705.2556	arxiv/0705.2496	arxiv/0705.2406
arxiv/0705.2258	arxiv/0705.2226	arxiv/0705.2171	arxiv/0705.2037	arxiv/0705.1920	arxiv/0705.1756
arxiv/0705.1720	arxiv/0705.1109	arxiv/0705.0934	arxiv/0705.0921	arxiv/0705.0579	arxiv/0705.0542
arxiv/0705.0521	arxiv/0705.0478	arxiv/0705.0153	arxiv/0705.0001	arxiv/0704.3925	arxiv/0704.3629
arxiv/0704.3543	arxiv/0704.3518	arxiv/0704.3285	arxiv/0704.3078	arxiv/0704.3064	arxiv/0704.2909
arxiv/0704.2816	arxiv/0704.2738	arxiv/0704.2595	arxiv/0704.2558	arxiv/0704.2543	arxiv/0704.2405
arxiv/0704.2276	arxiv/0704.2037	arxiv/0704.1999	arxiv/0704.1658	arxiv/0704.1590	arxiv/0704.1324
arxiv/0704.1044	arxiv/0704.0944	arxiv/0704.0794	arxiv/0704.0740	arxiv/0704.0674	arxiv/0704.0510
arxiv/0704.0371	arxiv/0704.0261	arxiv/0704.0222	arxiv/0704.0003		

† This is a `docx` file and was not processed.

‡ This paper was included in the search results, but was withdrawn and is not included in the Dark Matter corpus for lack of text. `gr-qc/0701040` was withdrawn for plagiarism.

Glossary

acyclic A network that does not contain cycles. 22, 104

adjacency matrix An $N \times N$ matrix representation of a network where N is the number of nodes in the network and the element $a_{ij} = 1$ if there is a link from node i to node j meaning that the nodes are adjacent in the network. 38, 74, 165

ADS Astrophysical Data Service. 13, 29, 30, 40, 56, 79, 82, 84, 98, 102, 118, 125, 128, 135–137, 163, 165, 167

betweenness centrality The number of shortest paths between all connected nodes that go through a link or node. 33, 38

bibcode Name for the identifiers used by ADS and described in Appendix A.1. 13, 30, 105

branch one of many subfields that make up a field. e.g. astrophysics is a branch of physics, computer security is a branch of computer science. 169

brane cosmology Related to string theory, it proposes that dark matter is a manifestation of matter existing in higher dimensions than our normal three dimensions of space. 99, 144

citation a citation is a reference in a scientific article to the source of the assertion, data or argument. They can be published in a scholarly journal, online on a pre-print server or indicate an unpublished source. In the terms of this thesis, a *citation* indicates a reference **from** another article and a *reference* is a reference **to** another article. 22

clique a subset of nodes that have links to each other node in the subset. The induced subgraph is complete . 23

clustering coefficient Defined in Equation 3.1, the clustering coefficient, C_i, is the measure of the number of neighbouring nodes which share links compared to the maximum possible number of nodes. The average clustering coefficient, $\bar{C}$,

is the average value of the clustering coefficient across all nodes in the network ($\bar{C} = \frac{1}{N}\sum_i^N C_i$). 25, 27

community While no standard definition exists, for the purposes of this thesis a community shall refer to a subset of nodes in a network that have a higher density of links from nodes in the subset to other nodes in the subset than to nodes not in the subset. See also the description of modularity, Section 3.4. When referring to scientific or research communities which indicate groups of people that are involved in research on related subjects, the relevant adjective will always be used. In contrast to unsupervised learning which produces clusters of objects that share similar properties, communities are formed by their structural connections. It is the links that tie communities together. 21, 33, 36, 38, 39, 88, 98, 115, 117, 119

complete In graph theory, a network in which every node is connected by a unique link. 33, 163

complex network As described in Section 3.1.1, the three characteristics that define a complex network are the small-world properties of random graphs, the tendency for links to cluster together (the clustering coefficient, C is much higher than the clustering coefficient of a random graph of similar size, $C \gg C_{rg}$) and is scale-free (the degree distribution has the fat tail typical of $P(k) \sim k^{-\gamma}$ for $\gamma > 0$). It has non-trivial topological features, not seen in lattices or random graphs. The difference between complex and merely complicated is that it possesses structural features on all scales as the result of spontaneous interaction among many individuals and is representative of the emergent behaviour of self-organising systems. 21, 24, 25, 27, 34, 42, 93, 103, 120

crisp set A set in which the members do not belong to other sets (the boundary is crisp or sharp), as opposed to a fuzzy set in which the set boundary is fuzzy and members near the boundary can belong to more than one set. 37

cycle In network theory, a path of links and nodes that ends up at the node it started from. Some types of cycle specify that no repetitions of links or nodes shall be used other than the starting node. 163, 165

Dark Matter corpus Depending on the stage of processing, it refers to the 2671 papers received from arXiv, the 2661 papers curated into the corpus having removed 8 papers from the `math` category and 2 papers removed from arXiv (listed in Table F.5) and the 2659 papers processed into VSM representation for text mining. 9, 12, 14, 16, 61, 63, 69, 71, 84, 93, 154, 165

Dark Matter network consists of the 778 491 nodes and 7 606 982 links that make up the core network and shells 1 and 2. It is built from the citation data found in the ADS. Diagrams of the structure of the network are found in Figures 2.4 and 2.5.

- **core network** consists of 2663 nodes representing the 2671 papers of the Dark Matter corpus with the 8 papers in the `math` category removed. 14, 26
- **full network** is the term used when contrasting the entire Dark Matter network with the core network or shells surrounding it. 26
- **shell** A shell consists of the citations and references found in the first or second pass of ADS queries. Shell 1 contains the 46 904 nodes that are references or citations to the core network and shell 2 contains the 728 926 nodes that are references or citations to shell 1. 14, 26

data mining The automated or semi-automated process of finding non-trivial and meaningful patterns in data. 48

degree (k) The number of links attached to a node.

- **in-degree** (k_{in}) The number of links arriving at a node in a directed network. ix, 22, 23, 31, 42–44
- **out-degree** (k_{out}) The number of links leaving a node in a directed network. 22, 23, 31, 42

diachronic Concerning the analysis of linguistic features as they change over time. 112, 114

directed acyclic graph (DAG) A directed network which does not contain cycles. 23, 41, 104

directed network A network in which the links retain the direction of travel *from* node i *to* node j. In the adjacency matrix, e_{ij} does not necessarily equal e_{ji}.

25, 26, 165

false friend Words that look or sound alike in two languages, but have significantly different meanings. 3

fat-tailed A probability distribution that exhibits larger values away from the central peak than would be expected in a gaussian or exponential distribution. Sometimes used interchangeably with *long-tailed distribution* by virtue of both being sub-classes of the heavy-tailed distributions in cases where the precise rate of decay is not under discussion. Unlike the gaussian distribution, the average value of the system is not typical. 24, 26, 28, 121

field An academic discipline which works within a commonly accepted methodology. e.g. physics and computer science are academic fields of study. 163

graph In mathematics, graph is a topological expression of relationships (called links) between entities (called nodes). In this thesis, a graph is equivalent to a network. It can have a simple structure like a lattice or grid, no coherent structure as in random graphs or non-trivial structures such as complex networks. 19

ground-truth Refers to information provided by direct observation as opposed to information provided by inference. 97, 102, 117

hapax legomenon A term that only appears once in a corpus or text document. 60

hub A highly-connected node. A node with significantly more links than the average node.. 21

interquartile range The range of numerical values stretching from the first quartile to the third quartile, where a quartile is one quarter of the data points. The interquartile range encompasses half of the measurements either side of the median value. 112

ISI Institute for Scientific Information. 24, 105, 170

jargon Terms used by professionals in a field that are inaccessible to non-experts. 3

keyword As used in Information Retrieval, a keyword is an index term that has significant meaning to the topic. Although used in the singular, it is not necessarily restricted to single words. Sometimes longer phrases may be called *key-phrases*. 56, 62, 67, 82, 89, 98, 112, 118

kurtosis From statistics, describes the higher-order, symmetrical deviation from the normal distribution. It is the fourth moment abut the mean, $\frac{\mu_4}{\sigma^2}$, where $\mu_4 = \frac{1}{n}\sum_i^n (x_i - \mu)^4$. A platykurtic curve is broader and flatter, whereas a leptokurtic curve is sharper and pointy. 67

Latent Semantic Indexing A technique in Text Mining of dimension reduction, based on singular value decomposition, represents documents in a conceptual space, overcoming issues of polysemy and synonymy. 49–51

LDA Latent Dirichlet Allocation. 61, 106

link Also called an edge or an arc, a link in a network is the topological representation of the relationship between entities. An example of a link is the citation between scientific articles. 19

Modified Newtonian Dynamics A model of new physics that alters the equations of Newton and Einstein at distances $> 10^{18}m$. 9, 167

MOND MOdified Newtonian Dynamics. 9, 29, 36, 82, 84, 87, 96, 100, 113, 120, 143, 144, *Glossary:* Modified Newtonian Dynamics

network A network is equivalent to a graph in this thesis. Two examples of networks are presented here, the Dark Matter citation network built from citation data from ADS and a document network created from applying a threshold to the text cosine similarity values (described in Section 4.8). In formal notation, given a set of nodes, $\mathbf{N}$, and a set of links, $\mathbf{L}$, a network $\mathbf{G}$ is the ordered pair, $\mathbf{G} = (\mathbf{N}, \mathbf{L})$. 22

neutrino (ν) Almost massless uncharged particle which only interacts via the weak force. 143

NLP Natural Language Processing. 49, 81, 124

node Also called a vertex, a node in a network is the topological representation of an entity. In this thesis, most nodes refer to documents. 19

PACS The Physics and Astronomy Classification Scheme is used by the journal *Physical Review* to identify topics in physics in the place of keywords yielding a standardised hierarchical description of the field. In the future, it may be replaced by a new scheme called Physics Subject Headings or PhySH. 112

path a sequence of links between connected nodes that describe a route traversed by an individual through a network from one node to another. In a directed graph, a path follows the direction of the links. 19, 22

polysemy The property of a word having multiple meanings. 114, 141

power law A term that describes a distribution that follows the form $y \propto x^n$ over at least five orders of magnitude. Its linear slope on a log-log plot is a defining feature, but is not unique to the distribution. Zipf's law is an example of this long-tailed distribution. It has no typical value and is one of the characteristics of scale-free networks. 21, 23, 24, 26, 31, 94, 110, 121, 122

pre-print A version of a scholarly article prior to the officially published article. These are used in fast-moving fields where quickly sharing experimental results is desired with the understanding that the editorial process of a peer-reviewed journal can take many months. 8, 13, 23, 104, 105, 128, 129

preferential attachment A process by which nodes with many links have a higher probability of attracting new links than nodes with fewer links. Known as cumulative advantage by Garfield. An example would be a model that chooses to add a new link to node i in a network with N links with probability $P(i) \propto k_i/N$. 21, 23, 45

random graph Also known as the Erdős-Rényi model, it is a graph or network generated by randomly linking nodes in the graph with equal probability. It has the small-world property that the mean shortest path length between two nodes of the graph grows in proportion to the logarithm of the number of nodes, $\langle l \rangle \propto log(N)$. 20, 24, 27, 96, 107

redshift see Appendix E. 99

residual sum of squares ($\widehat{RSS}_{min}$) The sum of the square of differences between the data y and the model $f(x)$ given by $\sum_{i=1}^{n}(y_i - f(x_i))^2$. 70, 71

scale-free A property that is independent of scale. There is no characteristic scale for the system which similar whether zoomed-in or zoomed-out. 24, 27, 41, 168

SMART notation In Information Retrieval, a scheme for describing the tf-idf variants for scoring VSM similarity between a document and a query. It compactly lists the weighting strategies for term frequency, document frequency and normalisation functions. See [Manning et al., 2008, Section 6.4]. 59

social network analysis The use of network concepts to study social interactions. Using people as nodes and their interactions as the links, sociologists have studied friendship networks, collaboration networks, communities on social media, disease transmission and sexual relationships. 21, 38

Standard Model The current best understanding of particle physics with relation to the strong and weak nuclear forces and the electromagnetic force. The Large Hadron Collider was built to test the Standard Model against newer theories. 144

sterile neutrino A neutrino which only interacts via the gravitational force. 143

subgraph A subset of nodes and links in a network. The subset of nodes must include all nodes on both ends of all the links included in the subset of links. In formal notation, a subgraph, $\mathbf{S}$, of the graph, $\mathbf{G} = (\mathbf{N}, \mathbf{L})$, is the graph $\mathbf{S} = (\mathbf{N_s}, \mathbf{L_s})$, where $\mathbf{N_s}$ is a subset of the nodes $\mathbf{N}$ (or $\mathbf{N_s} \subset \mathbf{N}$) and $\mathbf{N_s}$ is a subset of the nodes $\mathbf{N}$ (or $\mathbf{L_s} \subset \mathbf{L}$). 33, 163

subject The general content under discussion. Can contain a number of topics. Not to be confused with the grammatical or linguistic sense where it refers to the subject of a sentence. In this thesis it can refer to the inquiry into the nature of Dark Matter in the branch of astrophysics or developments in cryptography in the branch of computer security. 4

supervised learning Algorithms for machine learning that operate with predefined

labels. This method is trained to apply labels to items with a given set of attributes on a dataset for which the labels are known. It is then tested for accuracy against a second known dataset. When the classifier has been trained, it can then be used to apply labels to previously unseen items. 81, 98

synonymy The property of two words having the same or similar meanings. 81

text mining The automated or semi-automated process of finding non-trivial and meaningful patterns in text. As text is generally unstructured, it usually requires pre-processing before the 'mining' can take place. 48

token A character sequence considered a meaningful element for analysis and passed on to future processing, loosely correlated to a "word". The process of tokenisation is described in Section 4.3.1. 17, 55, 57

Topic Modelling A technique in Text Mining of dimension reduction which statistically analyses document collections to discover the latent topics contained in the collection, described in Section 6.1.2. 50, 51, 106, 126

unsupervised learning Algorithms for machine learning that operate without predefined labels. This method attempts to cluster similar items together. It can be used to discover unanticipated patterns in the data. 35, 51, 53, 54, 58, 59, 61, 63, 66, 70, 77, 79, 81, 92, 122, 125, 164

Vector Space Model A representation of text that counts frequencies of terms in a document or document fragment without regard to their order or position in the text. 18, 50, 103, 170

VSM Vector Space Model. 18, 50, 58, 59, 62–64, 71, 91, 103, 113, *Glossary:* Vector Space Model

Web of Science Owned by Thomson Reuters, previously known as the ISI. 10, 170

WoS Web of Science. 10, 11, *Glossary:* Web of Science

Zipf's law describes the frequencies of terms in a collection of documents noting that the frequency is inversely proportional to its rank, $cf_i \propto \frac{1}{i}$ where t_1 is the

most common term in the collection, t_2 the second most common, and so on [Manning et al., 2008, p 82]. 72, 114

References

Steven Abney. *Semisupervised Learning for Computational Linguistics*. CRC Press, Sep 2007.

Helmut A. Abt. Long-Term Citation Histories of Astronomical Papers. *Publications of the Astronomical Society of the Pacific*, 93:207, Apr 1981.

Helmut A. Abt. Citations to Single and Multiauthored Papers. *Publications of the Astronomical Society of the Pacific*, 96:746, Sep 1984.

Helmut A. Abt. Reference Frequencies in Astronomy and Related Sciences. *Publications of the Astronomical Society of the Pacific*, 99:1329, Dec 1987.

Helmut A. Abt. What Fraction of Literature References Are Incorrect? *Publications of the Astronomical Society of the Pacific*, 104:235, Mar 1992.

Helmut A. Abt. How Long Are Astronomical Papers Remembered? *Publications of the Astronomical Society of the Pacific*, 108:1059–1061, Oct 1996.

Helmut A. Abt. Is the Astronomical Literature Still Expanding Exponentially? *Publications of the Astronomical Society of the Pacific*, 110:210–213, Feb 1998a.

Helmut A. Abt. Why some papers have long citation lifetimes. *Nature*, 395:756–757, Oct 1998b.

Helmut A. Abt. A Comparison of the Citation Counts in the Science Citation Index and the NASA Astrophysics Data System. *Organizations and Strategies in Astronomy, Vol. 7*, 335:169–174, Jan 2006.

Alberto Accomazzi and Rahul Dave. Semantic Interlinking of Resources in the Virtual Observatory Era. *Astronomical Data Analysis Software and Systems*, 442(XX), 2011.

Alberto Accomazzi, Carolyn S. Grant, Guenther Eichhorn, Michael J. Kurtz, and Stephen S. Murray. The World Wide Web and ADS Services. In *Bulletin of the American Astronomical Society*, volume 185, page 1370. American Astronomical Society, Dec 1994.

Alberto Accomazzi, Guenther Eichhorn, Michael J. Kurtz, Carolyn S. Grant, and Stephen S. Murray. The NASA Astrophysics Data System: Architecture. *Astronomy and Astrophysics Supplement Series*, 143:85–109, Apr 2000.

Charu C. Aggarwal and ChengXiang Zhai. An Introduction to Text Mining. In Charu Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*, pages 1–10. Springer US, 2012a.

Charu C. Aggarwal and ChengXiang Zhai. A Survey of Text Classification Algorithms. In Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*, pages 163–222. Springer US, 2012b. DOI: 10.1007/978-1-4614-3223-4_6.

Charu C. Aggarwal and ChengXiang Zhai. A Survey of Text Clustering Algorithms. In Charu Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*, pages 77–128. Springer US, 2012c.

Alasdair Allen. Astro::ADS - An object orientated interface to NASA's ADS database, 2003. URL `https://metacpan.org/pod/Astro::ADS`.

Ethem Alpaydin. *Introduction to machine learning*. adaptive computation and machine learning. MIT Press, 2004.

José Ignacio Alvarez-Hamelin, Luca Dall'Asta, Alain Barrat, and Alessandro Vespignani. k-core decomposition: a tool for the visualization of large scale networks. *arXiv:cs/0504107*, Apr 2005.

Sophia Ananiadou and John McNaught. *Text mining for biology and biomedicine*. Citeseer, 2006.

Ion Androutsopoulos, Graeme D Ritchie, and Peter Thanisch. Natural language interfaces to databases an introduction. *Natural language engineering*, 1(01):29–81, 1995.

Ralitsa Angelova and Stefan Siersdorfer. A Neighborhood-based Approach for Clustering of Linked Document Collections. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, CIKM '06, pages 778–779, New York, NY, USA, 2006. ACM.

Grigoris Antoniou and Frank van Harmelen. Web Ontology Language: OWL. In Steffen Staab and Rudi Studer, editors, *Handbook on Ontologies*, International Handbooks on Information Systems, pages 91–110. Springer Berlin Heidelberg, 2009.

arXiv.org. arXiv monthly submission rate statistics, 1 Jan '10, Jan 2010. URL `https://arxiv.org/help/stats/2009_by_area/index`.

arXiv.org. arXiv Primer, 2014. URL `http://arxiv.org/help/primer`.

R Baeza-Yates and B Ribeiro-Neto. *Modern Information Retrieval: The Concepts and Technology Behind Search*. Pearson Education, Harlow, England, 2nd edition, 2011.

Frank Ball, Denis Mollison, and Gianpaolo Scalia-Tomba. Epidemics with two levels of mixing. *The Annals of Applied Probability*, pages 46–89, 1997.

Albert-Laszló Barabási. *Linked; the new science of networks*. Perseus Publishing, Cambridge, Mass, 2002.

Albert-László Barabási and Réka Albert. Emergence of Scaling in Random Networks. *Science*, 286(5439):509 –512, Oct 1999.

S. Basu, A. Banerjee, and R. Mooney. Active Semi-Supervision for Pairwise Constrained Clustering. In *Proceedings of the 2004 SIAM International Conference on Data Mining*, Proceedings, pages 333–344. Society for Industrial and Applied Mathematics, Apr 2004.

Stefano Battiston, Michele Catanzaro, and Guido Caldarelli. Social and Financial Networks. In Guido Caldarelli and Alessandro Vespignani, editors, *Large scale structure and dynamics of complex networks: from information technology to finance and natural science*, volume 2 of *Complex Systems and Interdisciplinary Science*. World Scientific, 2007.

M. G. Beiró, J. I. Alvarez-Hamelin, and J. R. Busch. A low complexity visualization tool that helps to perform complex systems analysis. *New Journal of Physics*, 10 (12):125003, Dec 2008.

Richard Ernest Bellman. *Adaptive Control Processes*. Princeton Press, Princeton, NJ, 1961.

Elisa Bellotti. The social processes of production and validation of knowledge in particle physics: Preliminary theoretical and methodological observations. *Procedia - Social and Behavioral Sciences*, 10:148–159, 2011.

P. Berkhin, Jacob Kogan, Charles Nicholas, and Marc Teboulle. A Survey of Clustering Data Mining Techniques. In *Grouping Multidimensional Data*, pages 25–71. Springer Berlin Heidelberg, 2006.

Samyukta Bhupatiraju, Önder Nomaler, Giorgio Triulzi, and Bart Verspagen. Knowledge flows Analyzing the core literature of innovation, entrepreneurship and science and technology studies. *Research Policy*, 41(7):1205–1218, Sep 2012.

James Binney and Scott Tremaine. *Galactic dynamics*. Princeton University Press, Princeton, N.J, 2nd edition, 2008.

Deborah Blackman and Angela M. Benson. Overcoming knowledge stickiness in scientific knowledge transfer. *Public Understanding of Science*, 21(5):573–589, Jul 2012.

R. D. Blandford and R. Narayan. Cosmological Applications of Gravitational Lensing. *Annual Review of Astronomy and Astrophysics*, 30(1):311–358, 1992.

David M Blei and John D Lafferty. A correlated topic model of science. *The Annals of Applied Statistics*, pages 17–35, 2007.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar 2003.

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, Oct 2008.

Mary L. Boas. *Mathematical methods in the physical sciences*. Wiley, Apr 1983.

Marián Boguñá, Romualdo Pastor-Satorras, and Alessandro Vespignani. Absence of Epidemic Threshold in Scale-Free Networks with Degree Correlations. *Physical Review Letters*, 90(2):028701, Jan 2003.

Katy Börner, Chaomei Chen, and Kevin W. Boyack. Visualizing knowledge domains. *Annual Review of Information Science and Technology*, 37(1):179–255, Jan 2003.

John A. Bullinaria and Joseph P. Levy. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39(3): 510–526, Aug 2007.

Joan L. Bybee. Diachronic Linguistics. *The Oxford Handbook of Cognitive Linguistics*, Jun 2010.

Guido Caldarelli and Alessandro Vespignani. *Large scale structure and dynamics of complex networks: from information technology to finance and natural science*. World Scientific, 2007.

Robert Callan. *Artificial intelligence*. Palgrave Macmillan, Basingstoke, 2003.

R. J. G. B. Campello. A fuzzy extension of the Rand index and other related indexes for clustering and classification assessment. *Pattern Recognition Letters*, 28(7):833–841, May 2007.

Bernard Carr. Baryonic Dark Matter. *Annual Review of Astronomy and Astrophysics*, 32(1):531–590, 1994.

Ciro Cattuto, Vittorio Loreto, and Luciano Pietronero. Semiotic dynamics and collaborative tagging. *Proceedings of the National Academy of Sciences*, 104(5):1461–1464, Jan 2007.

Paolo Ceravolo, Angelo Corallo, Ernesto Damiani, Gianluca Elia, Marco Viviani, and Antonio Zilli. Bottom-up extraction and maintenance of ontology-based metadata. In Elie Sanchez, editor, *Capturing Intelligence*, volume 1 of *Fuzzy Logic and the Semantic Web*, pages 265–282. Elsevier, 2006.

Tanmoy Chakraborty, Niloy Ganguly, and Animesh Mukherjee. Rising popularity of interdisciplinary research - An analysis of citation networks. In *COMSNETS*, pages 1–6, 2014.

Jonathan Chang and David M Blei. Relational topic models for document networks. In *AIStats*, volume 9, pages 81–88, Florida, USA, 2009. JMLR.

S. Chang, C. C. Aggarwal, and T. S. Huang. Learning Local Semantic Distances with Limited Supervision. In *2014 IEEE International Conference on Data Mining (ICDM)*, pages 70–79, Dec 2014.

Chaomei Chen. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for information Science and Technology*, 57(3):359–377, 2006.

Chaomei Chen, Fidelia Ibekwe-SanJuan, and Jianhua Hou. The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis. *Journal of the American Society for Information Science and Technology*, 61(7):1386–1409, 2010. arXiv:1002.1985.

P. Chen and S. Redner. Community structure of the physical review citation network. *Journal of Informetrics*, 4(3):278–290, Jul 2010.

Jinho Choi, Sangyoon Yi, and Kun Chang Lee. Analysis of keyword networks in MIS research and implications for predicting knowledge evolution. *Information & Management*, 48(8):371–381, Dec 2011.

A. K. Choudhary, P. I. Oluikpe, J. A. Harding, and P. M. Carrillo. The needs and benefits of Text Mining applications on Post-Project Reviews. *Computers in Industry*, 60(9):728–740, Dec 2009.

Aaron Clauset, M. E. J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical Review E*, 70(6):066111, Dec 2004.

Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *0706.1062*, Jun 2007. SIAM Review 51, 661-703 (2009).

Pol Colomer-de Simón and Marián Boguñá. Double percolation phase transition in clustered complex networks. *arXiv:1401.8176 [cond-mat, physics:physics]*, Jan 2014.

Luciano da Fontoura Costa, Osvaldo N. Oliveira, Gonzalo Travieso, Francisco Aparecido Rodrigues, Paulino Ribeiro Villas Boas, Lucas Antiqueira, Matheus Palhares Viana, and Luis Enrique Correa Rocha. Analyzing and modeling real-world phenomena with complex networks: a survey of applications. *Advances in Physics*, 60 (3):329–412, May 2011.

Steven P. Crain, Ke Zhou, Shuang-Hong Yang, and Hongyuan Zha. Dimensionality Reduction and Topic Modeling: From Latent Semantic Indexing to Latent Dirichlet Allocation and Beyond. In Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*, pages 129–161. Springer US, 2012. DOI: 10.1007/978-1-4614-3223-4_5.

W Bruce Croft, Donald Metzler, and Trevor Strohman. *Search engines: Information retrieval in practice*, volume 283. Addison-Wesley Reading, 2010.

Ryan R Curtin. Dual-tree k-means with bounded iteration runtime. *arXiv preprint arXiv:1601.03754*, 2016.

Douglass R. Cutting, David R. Karger, Jan O. Pedersen, and John W. Tukey. Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '92, pages 318–329, New York, NY, USA, 1992. ACM.

Walter Daelemans and Antal van den Bosch. *Memory-Based Language Processing*. Cambridge University Press, Sep 2005.

John Daintith and R. D. Nelson. *The Penguin Dictionary of Mathematics*. Penguin Books Ltd., London, reprint edition, Apr 1989.

Leon Danon, Jordi Duch, Alex Arenas, and Albert Díaz-Guilera. Community Structure Identification. In Guido Caldarelli and Alessandro Vespignani, editors, *Large scale structure and dynamics of complex networks: from information technology to finance and natural science*, volume 2 of *Complex Systems and Interdisciplinary Science*. World Scientific, 2007.

D. L. Davies and D. W. Bouldin. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, Apr 1979.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391, 1990.

Dursun Delen and Martin D. Crossland. Seeding the survey and analysis of research literature with text mining. *Expert Systems with Applications*, 34(3):1707–1720, Apr 2008.

Inderjit S Dhillon and Dharmendra S Modha. Concept decompositions for large sparse text data using clustering. *Machine learning*, 42(1-2):143–175, 2001.

Angelo Di Iorio, Andrea Giovanni Nuzzolese, and Silvio Peroni. Towards the Automatic

Identification of the Nature of Citations. In *Proceedings of 3rd Workshop on Semantic Publishing*, pages 63–74, Aachen, Germany, 2013.

Serge N. Dorogovtsev and José F. F. Mendes. *Evolution of networks: from biological nets to the Internet and WWW*. Oxford University Press, 2003.

Richard Duda, Peter Hart, and David Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, Nov 2001.

R. I. M. Dunbar. Neocortex size as a constraint on group size in primates. *Journal of Human Evolution*, 22(6):469–493, Jun 1992.

J. C. Dunn. Well-Separated Clusters and Optimal Fuzzy Partitions. *Journal of Cybernetics*, 4(1):95–104, Jan 1974.

Guenther Eichhorn, Michael J. Kurtz, Alberto Accomazzi, Carolyn S. Grant, and Stephen S. Murray. The NASA Astrophysics Data System: The search engine and its user interface. *Astronomy and Astrophysics Supplement Series*, 143(1):61–83, Apr 2000.

Guenther Eichhorn, Alberto Accomazzi, Carolyn S. Grant, Michael J. Kurtz, Vicente Reybacaicoa, and Stephen S. Murray. The ADS Abstract Service: A Free NASA Service. In *APS April Meeting*, volume APR02, page 11007, Albuquerque, New Mexico, Apr 2002. American Physical Society.

Elena Erosheva, Stephen Fienberg, and John Lafferty. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5220–5227, Jun 2004.

T. S. Evans, R. Lambiotte, and P. Panzarasa. Community structure and patterns of scientific collaboration in Business and Management. *Scientometrics*, 89(1):381–396, Oct 2011.

Giorgio Fagiolo. Clustering in complex directed networks. *Phys. Rev. E*, 76:026107, Aug 2007.

Matthew E Falagas, Eleni I Pitsouni, George A Malietzis, and Georgios Pappas. Comparison of PubMed, Scopus, Web of Science, and Google scholar: strengths and weaknesses. *The FASEB journal*, 22(2):338–342, 2008.

Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37, 1996.

Jonathan L. Feng. Dark Matter Candidates from Particle Physics and Methods of Detection. *Annual Review of Astronomy and Astrophysics*, 48(1):495–545, 2010.

Rana Forsati, Mehrdad Mahdavi, Mehrnoush Shamsfard, and Mohammad Reza Meybodi. Efficient stochastic algorithms for document clustering. *Information Sciences*, 220:269–291, Jan 2013.

Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(35):75–174, Feb 2010.

Santo Fortunato and Darko Hric. Community detection in networks: A user guide. *Physics Reports*, 659:1–44, 2016.

Noah E. Friedkin. Information flow through strong and weak ties in intraorganizational social networks. *Social Networks*, 3(4):273–285, 1982.

M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):7821 –7826, Jun 2002.

Eric J. Glover, Kostas Tsioutsiouliklis, Steve Lawrence, David M. Pennock, and Gary W. Flake. Using Web Structure for Classifying and Describing Web Pages. In *Proceedings of the 11th International Conference on World Wide Web*, WWW '02, pages 562–569, New York, NY, USA, 2002. ACM.

Michael Golosovsky and Sorin Solomon. The Transition Towards Immortality: Non-linear Autocatalytic Growth of Citations to Scientific Papers. *Journal of Statistical Physics*, 151(1-2):340–354, Apr 2013.

Michael Golosovsky and Sorin Solomon. Uncovering the dynamics of citations of scientific papers. *arXiv:1410.0343 [physics]*, Oct 2014. arXiv: 1410.0343.

Michael Golosovsky and Sorin Solomon. Growing complex network of citations of scientific papers: Modeling and measurements. *Physical Review E*, 95(1):012324, Jan 2017.

Bruno Goncalves, Nicola Perra, and Alessandro Vespignani. Validation of Dunbar's number in Twitter conversations. *PLoS ONE*, 6(8):e22656, Aug 2011.

Mark S. Granovetter. The Strength of Weak Ties. *American Journal of Sociology*, 78 (6):1360–1380, May 1973.

Carolyn S. Grant, Alberto Accomazzi, Guenther Eichhorn, Michael J. Kurtz, and Stephen S. Murray. The NASA Astrophysics Data System: Data holdings. *Astronomy and Astrophysics Supplement Series*, 143(1):111–135, Apr 2000.

John Gribbin. *The stuff of the Universe: dark matter, mankind and the coincidences of cosmology*. Heinemann, London, 1990.

H. P. Grice. *Studies in the way of words*. Harvard University Press, Cambridge, Mass, 1989.

Georg Groh and Christoph Fuchs. Multi-modal social networks for modeling scientific fields. *Scientometrics*, 2011.

Marco Guerini, Alberto Pepe, and Bruno Lepri. Do Linguistic Style and Readability of Scientific Abstracts Affect their Virality? In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

Himanshu Gupta and Rajeev Srivastava. k-means Based Document Clustering with Automatic k Selection and Cluster Refinement. *International Journal of Computer Science and Mobile Applications*, 2(5):7–13, 2014.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*, 2016.

Youngsub Han and Yanggon Kim. A Method of Discovering Genre Similarity Using Aspect Based Approach. In *Proceedings of the International Conference on Research in Adaptive and Convergent Systems*, RACS '16, pages 29–34, New York, NY, USA, 2016. ACM.

David J Hand, Heikki Mannila, and Padhraic Smyth. *Principles of data mining*. MIT press, 2001.

Mary Dee Harris. *Introduction to Natural Language Processing*. Reston Publishing Co., Reston, VA, USA, 1985.

Zellig S. Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

Hussein Hashimi, Alaaeldin Hafez, and Hassan Mathkour. Selection criteria for text mining approaches. *Computers in Human Behavior*, 51, Part B:729–733, Oct 2015.

Marti A. Hearst. Untangling Text Data Mining. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 3–10, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics.

Thomas Hofmann. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 50–57, New York, NY, USA, 1999. ACM.

A Hotho, S. Staab, and G. Stumme. Ontologies improve text document clustering. In *Third IEEE International Conference on Data Mining, 2003. ICDM 2003*, pages 541–544, Nov 2003.

Darko Hric, Richard K. Darst, and Santo Fortunato. Community detection in networks: structural clusters versus ground truth. *arXiv:1406.0146 [physics, q-bio]*, Jun 2014.

Faliang Huang, Shichao Zhang, Minghua He, and Xindong Wu. Clustering web documents using hierarchical representation with multi-granularity. *World Wide Web*, 17(1):105–126, Jan 2014.

Kejun Huang, Xiao Fu, and Nicholas D. Sidiropoulos. Anchor-Free Correlated Topic Modeling: Identifiability and Algorithm. *arXiv:1611.05010 [cs, stat]*, Nov 2016. arXiv: 1611.05010.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, Dec 1985.

Ioana Hulpus, Conor Hayes, Marcel Karnstedt, and Derek Greene. Unsupervised Graph-based Topic Labelling Using DBpedia. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 465–474, New York, NY, USA, 2013. ACM.

Ammerah Jabr-Hamdan, Jie Sun, and Daniel ben Avraham. Growing Networks with Super-Joiners. *arXiv:1405.7018 [cond-mat, physics:physics]*, May 2014.

Anil K. Jain. Data clustering: 50 years beyond K-means. *Award winning papers from the 19th International Conference on Pattern Recognition (ICPR)19th International Conference in Pattern Recognition (ICPR)*, 31(8):651–666, Jun 2010.

Anil K. Jain and Richard C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.

Shivanjli Jain and Amanjot Kaur Grewal. Comparison of Clustering Algorithms Based on Outliers. *International Research Journal of Engineering and Technology*, 3(11), Nov 2016.

James Jardine and Simone Teufel. Topical PageRank: A Model of Scientific Expertise for Bibliographic Search. In *14th Conference of the European Chapter of the ACL*, pages 501–510, Gothenburg, Sweden, Apr 2014. Association for Computational Linguistics.

N. Jardine and C.J. van Rijsbergen. The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval*, 7(5):217–240, Dec 1971.

Jing Jiang. Information Extraction from Text. In Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*, pages 11–41. Springer US, 2012. DOI: 10.1007/978-1-4614-3223-4_2.

Yookyung Jo, John E. Hopcroft, and Carl Lagoze. The Web of Topics: Discovering the Topology of Topic Evolution in a Corpus. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 257–266, New York, NY, USA, 2011. ACM.

Karen Sparck Jones. *Readings in information retrieval*. Morgan Kaufmann, 1997.

Daniel Jurafsky and James Martin. *Speech and language processing: An introduction to speech recognition*. Pearson Prentice Hall, 2nd edition, 2008.

Daniel Kahneman, Paul Slovic, and Amos Tversky. *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press, Apr 1982.

Brian Karrer and M. E. J. Newman. Random Acyclic Networks. *Physical Review Letters*, 102(12):128701, Mar 2009.

Márton Karsai, Nicola Perra, and Alessandro Vespignani. Time varying networks and the weakness of strong ties. *Scientific Reports*, 4, Feb 2014.

Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, Sep 2009.

Sukhpal Kaur and Er Mamoon Rashid. Web News Mining using Back Propagation Neural Network and Clustering using K-Means Algorithm in Big Data. *Indian Journal of Science and Technology*, 9(41), Nov 2016.

Meen Chul Kim and Chaomei Chen. A scientometric review of emerging trends and new developments in recommendation systems. *Scientometrics*, 104(1):239–263, Jul 2015.

Lawrence Maxwell Krauss. *Quintessence; the mystery of missing mass in the universe.* Vintage, London, 2001.

Thomas S. Kuhn. *The Structure of Scientific Revolutions.* University of Chicago Press, Chicago, 1962.

Michael J. Kurtz, Guenther Eichhorn, Alberto Accomazzi, Carolyn S. Grant, Stephen S. Murray, and Joyce M. Watson. The NASA Astrophysics Data System: Overview. *Astronomy and Astrophysics Supplement Series*, 143(1):41–59, Apr 2000.

Thomas K Landauer. On the computational basis of learning and cognition: Arguments from LSA. In Brian H. Ross, editor, *Psychology of Learning and Motivation*, volume Volume 41, pages 43–84. Academic Press, 2002.

Thomas K. Landauer and Susan T. Dumais. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240, 1997.

Jean Lave. *Situated learning: legitimate peripheral participation.* Cambridge University Press, Cambridge, 1991.

Conrad Lee and Pádraig Cunningham. Community detection: effective evaluation on large social networks. *Journal of Complex Networks*, 2(1):19–37, Jan 2014.

M. Lee, B. Pincombe, and M. Welsh. An empirical evaluation of models of text document similarity. In *XXVII Annual Conference of the Cognitive Science Society*, Stresa, Italy, 2005. Cognitive Science Society.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. DBpedia A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2):167–195, Jan 2015.

Fredrik Liljeros, Christofer R. Edling, Luis A. Nunes Amaral, H. Eugene Stanley, and Yvonne Aberg. The web of human sexual contacts. *Nature*, 411(6840):907–908, Jun 2001.

Kar Wai Lim and Wray L Buntine. Bibliographic Analysis with the Citation Network Topic Model. In *Proceedings of the 6th Asian Conference on Machine Learning*, pages 142–158. JMLR, 2014.

Lucas A. Lopes, Vinicius P. Machado, Ricardo A. L. Rabêlo, Ricardo A. S. Fernandes, and Bruno V. A. Lima. Automatic labelling of clusters of discrete and continuous data with supervised machine learning. *Knowledge-Based Systems*, 106:231–241, Aug 2016.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. Cambridge University Press, 2008.

Daniel A. McFarland, Daniel Ramage, Jason Chuang, Jeffrey Heer, Christopher D. Manning, and Daniel Jurafsky. Differentiating language usage through topic models. *Poetics*, 41(6):607–625, Dec 2013.

Deborah L. McGuinness. Owl web ontology language reference, Feb 2004. URL `http://www.academia.edu/2805156/Owl_web_ontology_language_reference`.

Qiaozhu Mei, Deng Cai, Duo Zhang, and ChengXiang Zhai. Topic Modeling with Network Regularization. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 101–110, New York, NY, USA, 2008. ACM.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*, Jan 2013. arXiv: 1301.3781.

Stanley Milgram. The Small World Problem. *Psychology Today*, 1:61–67, May 1967.

Bruce Miller. LaTeXML:The Manual, 2014. URL `http://dlmf.nist.gov/LaTeXML/manual/`.

R. K. Mishra, K. Saini, and S. Bagri. Text document clustering on the basis of inter passage approach by using K-means. In *2015 International Conference on Computing, Communication Automation (ICCCA)*, pages 110–113, May 2015.

Richard Montague. Pragmatics. In RH Thomason, editor, *Formal Philosophy*, pages 95–118. Yale University Press, 1974.

Gheorghe Muresan and David J. Harper. Topic modeling for mediated access to very large document collections. *Journal of the American Society for Information Science and Technology*, 55(10):892–910, 2004.

Rahul Nalawade, Akash Samal, and Kiran Avhad. Improved Similarity Measure For Text Classification And Clustering. *International Research Journal of Engineering and Technology*, 3(05):214–219, 2016.

L. F. d C. Nassif and E. R. Hruschka. Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection. *IEEE Transactions on Information Forensics and Security*, 8(1):46–54, Jan 2013.

Richard E. Neapolitan. *Foundations of Algorithms: Using C++ Pseudocode*. Jones and Bartlett Publishers, Inc, Sudbury, Mass, 3rd edition, Jul 2003.

Ani Nenkova and Kathleen McKeown. A Survey of Text Summarization Techniques. In Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*, pages 43–76. Springer US, 2012. DOI: 10.1007/978-1-4614-3223-4_3.

M. E. J. Newman. Analysis of weighted networks. *Physical Review E*, 70(5):056131, Nov 2004.

M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64(2):026118, Jul 2001.

Mark E. J. Newman, Albert-Laszló Barabási, and Duncan J. Watts. *The structure and dynamics of networks*. Princeton University Press, Apr 2006.

Robert Nisbet, John Elder, and Gary Miner. *Handbook of Statistical Analysis and Data Mining Applications*. Academic Press, 2009.

Jesús Oliva, José Ignacio Serrano, María Dolores del Castillo, and Ángel Iglesias. SyMSS: A syntax-based measure for short-text semantic similarity. *Data & Knowledge Engineering*, 70(4):390–405, Apr 2011.

Jeremiah P. Ostriker. Astronomical Tests of the Cold Dark Matter Scenario. *Annual Review of Astronomy and Astrophysics*, 31(1):689–716, 1993.

Romualdo Pastor-Satorras, Alexei Vázquez, and Alessandro Vespignani. Dynamical and Correlation Properties of the Internet. *Physical Review Letters*, 87(25):258701, Nov 2001. arXiv: cond-mat/0105161.

Deana L. Pennell and Yang Liu. Normalization of informal text. *Computer Speech & Language*, 28(1):256–277, Jan 2014.

Georg Peters and Richard Weber. DCC: a framework for dynamic granular clustering. *Granular Computing*, 1(1):1–11, 2016.

Ithiel de Sola Pool and Manfred Kochen. Contacts and influence. *Social Networks*, 1 (1):5–51, 1978.

Alexandrin Popescul and Lyle H Ungar. Automatic labeling of document clusters. *Unpublished manuscript, available at http://citeseer. nj. nec. com/popescul00automatic. html*, 2000.

Martin Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

Troy A. Porter, Robert P. Johnson, and Peter W. Graham. Dark Matter Searches with Astroparticle Data. *Annual Review of Astronomy and Astrophysics*, 49(1):155–194, 2011.

Derek J. de Solla Price. Networks of Scientific Papers. *Science*, 149(3683):510 –515, Jul 1965.

Derek J. de Solla Price. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5): 292–306, Sep 1976.

Mohammad Rabiei, Seyyed-Mahdi Hosseini-Motlagh, and Abdorrahman Haeri. Using text mining techniques for identifying research gaps and priorities: a case study of the environmental science in Iran. *Scientometrics*, pages 1–28, Dec 2016.

Dragomir R. Radev, Mark Hodges, Anthony Fader, Mark Joseph, Joshua Gerrish, Mark Schaller, Jonathan dePeri, and Bryan Gibson. Clairlib documentation v1.07. Technical Report CSE-TR-536-07, University of Michigan. Department of Electrical Engineering and Computer Science, 2007.

Filippo Radicchi, Claudio Castellano, Federico Cecconi, Vittorio Loreto, and Domenico Parisi. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9):2658–2663, Feb 2004.

Daniel Ramage, Christopher D. Manning, and Susan Dumais. Partially Labeled Topic Models for Interpretable Text Mining. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 457–465, New York, NY, USA, 2011. ACM.

Erzsébet Ravasz and Albert-László Barabási. Hierarchical organization in complex networks. *Physical Review E*, 67(2):026112, Feb 2003.

S. Redner. How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B - Condensed Matter and Complex Systems*, 4(2): 131–134, Jul 1998.

S. U. Rehman, S. Asghar, S. Fong, and S. Sarasvady. DBSCAN: Past, present and future. In *The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)*, pages 232–238, Feb 2014.

Jim Richardson. Lingua::Stem::En, Jun 1999. URL `https://metacpan.org/pod/Lingua::Stem::En`.

C.J. van Rijsbergen. Information retrieval: theory and practice. In *Proceedings of the Joint IBM/University of Newcastle upon Tyne Seminar on Data Base Systems*, pages 1–14, 1979.

Ian Robinson, Jim Webber, and Emil Eifrem. *Graph Databases*. O'Reilly Media, Sebastopol, Calif., 1st edition, Jun 2013.

Everett M. Rogers. *Diffusion of innovations*. Free Press, New York, 1962.

Yonatan Rosen and Yoram Louzoun. Topological similarity as a proxy to content similarity. *Journal of Complex Networks*, 4(1):38–60, Mar 2016.

Martin Rosvall and Carl T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105 (4):1118–1123, Jan 2008.

Vera C. Rubin and W. Kent Ford, Jr. Rotation of the Andromeda Nebula from a Spectroscopic Survey of Emission Regions. *The Astrophysical Journal*, 159:379, Feb 1970.

Gordon Rugg and Peter McGeorge. The sorting techniques: a tutorial paper on card sorts, picture sorts and item sorts. *Expert Systems*, 22(3):94–107, Jul 2005.

G. Salton, A. Wong, and C. S. Yang. A Vector Space Model for Automatic Indexing. *Commun. ACM*, 18(11):613–620, Nov 1975.

Robert H. Sanders and Stacy S. McGaugh. Modified Newtonian Dynamics as an Alternative to Dark Matter. *Annual Review of Astronomy and Astrophysics*, 40(1): 263–317, 2002.

Michael T. Schaub, Renaud Lambiotte, and Mauricio Barahona. Encoding dynamics for multiscale community detection: Markov time sweeping for the Map equation. *Physical Review E*, 86(2), Aug 2012.

Michael T. Schaub, Jean-Charles Delvenne, Martin Rosvall, and Renaud Lambiotte. The many facets of community detection in complex networks. *arXiv:1611.07769 [physics]*, Nov 2016. arXiv: 1611.07769.

Hinrich Schütze and Craig Silverstein. Projections for efficient document clustering. In *ACM SIGIR Forum*, volume 31, pages 74–81. ACM, 1997.

John Scott. *Social Network Analysis: A Handbook*. SAGE Publications, Jan 2000.

Toby Segaran. *Programming collective intelligence: building smart web 2.0 applications*. O'Reilly Media, Inc., 2007.

María Angeles Serrano, Marian Boguñá, Romualdo Pastor-Satorras, and Alessandro Vespignani. Correlations in complex networks. *Large scale structure and dynamics of complex networks: From information technology to finance and natural sciences*, pages 35–66, 2007.

David Shotton. CiTO, the Citation Typing Ontology. *Journal of Biomedical Semantics*, 1(Suppl 1):S6, 2010.

Joseph Silk. *The Big Bang*. W.H. Freeman, New York, 3rd edition, 2001.

M. V. Simkin and V. P. Roychowdhury. Copied citations create renowned papers? *Annals of Improbable Research*, 11(1):24–27, Jan 2005. arXiv: cond-mat/0305150.

M. V. Simkin and V. P. Roychowdhury. Re-inventing Willis. *Physics Reports*, 502(1): 1–35, May 2011.

Herbert A. Simon. On a Class of Skew Distribution Functions. *Biometrika*, 42(3/4): 425, Dec 1955.

Matthew S. Simpson and Dina Demner-Fushman. Biomedical Text Mining: A Survey of Recent Progress. In Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*, pages 465–517. Springer US, 2012. DOI: 10.1007/978-1-4614-3223-4_14.

André Skupin. Discrete and continuous conceptualizations of science: Implications for knowledge domain visualization. *Journal of Informetrics*, 3(3):233–245, Jul 2009.

Richard Sproat, Alan W. Black, Stanley Chen, Shankar Kumar, Mari Ostendorf, and Christopher Richards. Normalization of non-standard words. *Computer Speech & Language*, 15(3):287–333, Jul 2001.

Matthew Steen, Satoru Hayasaka, Karen Joyce, and Paul Laurienti. Assessing the consistency of community structure in complex networks. *Physical Review E*, 84(1), Jul 2011.

Benno Stein and Sven Meyer Zu Eissen. Topic Identification: Framework and Application. In *I-KNOW '04*, Graz, Austria, 2004.

Michael Steinbach, George Karypis, and Vipin Kumar. A comparison of document clustering techniques. In *KDD Workshop on Text Mining*, 2000.

Michael Steinbach, Levent Ertöz, and Vipin Kumar. The challenges of clustering high dimensional data. In *New Directions in Statistical Physics*, pages 273–309. Springer Berlin Heidelberg, 2004.

Lovro Šubelj, Nees Jan van Eck, and Ludo Waltman. Clustering Scientific Publications Based on Citation Relations: A Systematic Comparison of Different Methods. *PLOS ONE*, 11(4):e0154404, Apr 2016.

Xiaoling Sun, Jasleen Kaur, Staša Milojević, Alessandro Flammini, and Filippo Menczer. Social Dynamics of Science. *Scientific Reports*, 3, Jan 2013.

Jian Tang, Cheng Li, Ming Zhang, and Qiaozhu Mei. Less is More: Learning Prominent and Diverse Topics for Data Summarization. *arXiv:1611.09921 [cs]*, Nov 2016. arXiv: 1611.09921.

Simone Teufel, Jean Carletta, and Marc Moens. An Annotation Scheme for Discourse-level Argumentation in Research Articles. In *Proceedings of the Ninth Conference on*

European Chapter of the Association for Computational Linguistics, EACL '99, pages 110–117, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. An Annotation Scheme for Citation Function. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, SigDIAL '06, pages 80–87, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

Stephen Edelston Toulmin. *Human Understanding Vol 1.* Clarendon Press, Oxford, 1972.

Jeffrey Travers and Stanley Milgram. An Experimental Study of the Small World Problem. *Sociometry*, 32(4):425–443, Dec 1969.

Sam Tregar. *Writing Perl Modules for CPAN.* Apress, Aug 2002.

Virginia Trimble. Existence and Nature of Dark Matter in the Universe. *Annual Review of Astronomy and Astrophysics*, 25(1):425–472, 1987.

Neil deGrasse Tyson. *Origins; fourteen billion years of cosmic evolution.* W.W. Norton & Co, New York, 2004.

Theresa Velden, Shiyan Yan, and Carl Lagoze. Mapping the cognitive structure of astrophysics by infomap clustering of the citation network and topic affinity analysis. *Scientometrics*, pages 1–19, Feb 2017.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. *Journal of Machine Learning Research*, 11(Oct):2837–2854, 2010.

Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation Methods for Topic Models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1105–1112, New York, NY, USA, 2009. ACM.

Chong Wang and David M. Blei. Collaborative Topic Modeling for Recommending Scientific Articles. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 448–456, New York, NY, USA, 2011. ACM.

Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, Jun 1998.

Xing Wei and W. Bruce Croft. LDA-based Document Models for Ad-hoc Retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 178–185, New York, NY, USA, 2006. ACM.

Sholom M. Weiss, Nitin Indurkhya, and Tong Zhang. Finding Structure in a Document Collection. In *Fundamentals of Predictive Text Mining*, Texts in Computer Science, pages 91–112. Springer London, Jan 2010a.

Sholom M. Weiss, Nitin Indurkhya, and Tong Zhang. *Fundamentals of Predictive Text Mining*. Springer Publishing Company, Incorporated, 1st edition, 2010b.

Etienne Wenger. *Communities of practice: learning, meaning, and identity*. Learning in doing: social cognitive, and computational perspectives. Cambridge University Press, Cambridge, 1999.

Anthony Weston. *A rulebook for arguments*. Hackett Publishing Company, Indianapolis, 4th edition, 2009.

Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques, Third Edition*. Morgan Kaufmann, 3rd edition, Jan 2011.

G. Wu, H. Lin, E. Fu, and L. Wang. An Improved K-means Algorithm for Document Clustering. In *2015 International Conference on Computer Science and Mechanical Automation (CSMA)*, pages 65–69, Oct 2015.

Junjie Wu, Hui Xiong, and Jian Chen. Adapting the Right Measures for K-means Clustering. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 877–886, New York, NY, USA, 2009. ACM.

Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, and Dan Steinberg. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37, Dec 2007.

Yan Wu, Tom Z. J. Fu, and Dah Ming Chiu. Generalized preferential attachment considering aging. *Journal of Informetrics*, 8(3):650–658, Jul 2014.

Zhi-Xi Wu and Petter Holme. Modeling scientific-citation patterns and other triangle-rich acyclic networks. *Physical Review E*, 80(3):037101, Sep 2009. arXiv:0908.2615.

Zheng Xie, Zhenzheng Ouyang, Qi Liu, and Jianping Li. A geometric graph model for citation networks of exponentially growing scientific papers. *Physica A: Statistical Mechanics and its Applications*, 456:167–175, Aug 2016.

Jaewon Yang and Jure Leskovec. Defining and Evaluating Network Communities Based on Ground-truth. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, MDS '12, pages 3:1–3:8, New York, NY, USA, 2012. ACM.

Jaewon Yang and Jure Leskovec. Structure and Overlaps of Ground-Truth Communities in Networks. *ACM Trans. Intell. Syst. Technol.*, 5(2):26:1–26:35, Apr 2014.

Yang Yang, Yuxiao Dong, and Nitesh V. Chawla. Microscopic Evolution of Social Networks by Triad Position Profile. *arXiv:1310.1525 [physics]*, Oct 2013.

Li Yueping. Fast computation of modularity in agglomerative clustering methods for community discovery. *IJACT: International Journal of Advancements in Computing Technology*, 3(4):153–164, 2011.

W Zachary. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, 33(4):452–473, 1977.

Xiaodan Zhang, Xiaohua Hu, and Xiaohua Zhou. A Comparative Evaluation of Different Link Types on Enhancing Document Clustering. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 555–562, New York, NY, USA, 2008. ACM.

Shi Zhong. Efficient streaming text clustering. *Neural Networks*, 18(56):790–798, Jul 2005.

F. Zwicky. Die Rotverschiebung von extragalaktischen Nebeln. *Helvetica Physica Acta*, 6:110–127, 1933.