Exploring Informatics Research Landscape: Collaborations and Topic Similarity

Wei Ting Goh

Honours Project Report
BSc Artificial Intelligence and Computer Science
School of Informatics
University of Edinburgh

2018

Abstract

Using bibliometric data from *Edinburgh Research Explorer*, we presented the first study on Informatics collaboration network and similarity network created using topic modelling algorithms to discover like-minded Informatics researchers. Studying various properties of collaboration network, we showed that Informatics is a small-world network where researchers are separated by four degrees on average, and collaborations within Informatics have increased over the years. Using the principle of homophily, we investigated if the likelihood of collaboration between similar individuals is more than a random phenomenon, and found evidence that collaboration usually occurs within institute more than between institutes.

We postulate that common field of interests engendered in each institute influenced choice of collaborators. Extending that idea, we leverage textual data underlying the same collaboration network to realise similarity networks by representing researchers using topics of computer science. We found that by using collaboration network to estimate the size of our similarity network, topic based communities are discovered. We call this network a topic-similarity network, which potentially could allow researchers to find like-minded others to innovate and collaborate.

Dynamic representations of these networks are provided in our companion website (`https://goweiting.github.io/infnet`) for readers to explore networks underlying School of Informatics.

Acknowledgements

Thanks to my parents and friends who have supported and took care of me throughout the last leg of my academic journey. For I would remember nothing but the memories we had in Edinburgh.

Many thanks to my project supervisor - Dr Rik Sarkar - for not just providing dedicated hardware for computation, advice on software and technical issues, but also heartware throughout the course of this project. Your unwavering support, patience, and dedication to the project is commendable. From the bottom of my heart, ***Thank You***.

Table of Contents

Chapter 1

Introduction

Collaboration is fundamental in scientific research and continues to increase in frequency and importance over the years. Not only does it has potential to solve complex scientific problems, collaboration leads to spreading of information, ideas, and innovation - transcending geographical boundaries and cultures (Grossman, 2002; Newman, 2004). This is especially apparent in computer science where the Fourth Industrial Revolution requires computer scientist to participate in research that leverage on technological systems, creating multidisciplinary field of interests, e.g. computational chemistry, bioinformatics, and digital humanities.

While there exists numerous studies on collaboration network using publicly available bibliometric data, none was conducted for School of Informatics to the best of our knowledge. Our primary task is to elucidate collaboration patterns using methods of network analysis. To that end, we showed that existing infrastructure from the University - *Edinburgh Research Explorer* - provides a mean to create accurate collaboration networks. We further our study by investigating similarity networks based on topics modelled from collection of published papers and discovered a new perspective on collaboration relationships. We call this network a topic-similarity network.

A **scientific collaboration network** has all its nodes as scientists who have published a research paper. A pair of nodes are connected if two scientist have one or more joint publications where each may or may not include others. There are two ways we can represent these edges: 1) As a simple undirected graph where edges represent the existence of collaboration; 2) As weighted edges to codify the strength of ties. Often in network science, the collaboration network is a subject of interest because of the ease of access to good quality bibliometric data on the Web, and its usefulness for understanding productivity and different collaboration patterns between disciplines. Interest of collaboration network is not just limited to network scientist but also sociologists and work psychologists (Goffman, 1969; Franceschet, 2010, 2011; Newman, 2001c,b). The collaboration network is however limited if the dataset contains multiple 'solo-publications' or accurate identification of scientists in publications is futile, resulting in numerous isolated nodes. Network analysis tools that take into account these nodes do not provide an meaningful characterisation of the network.

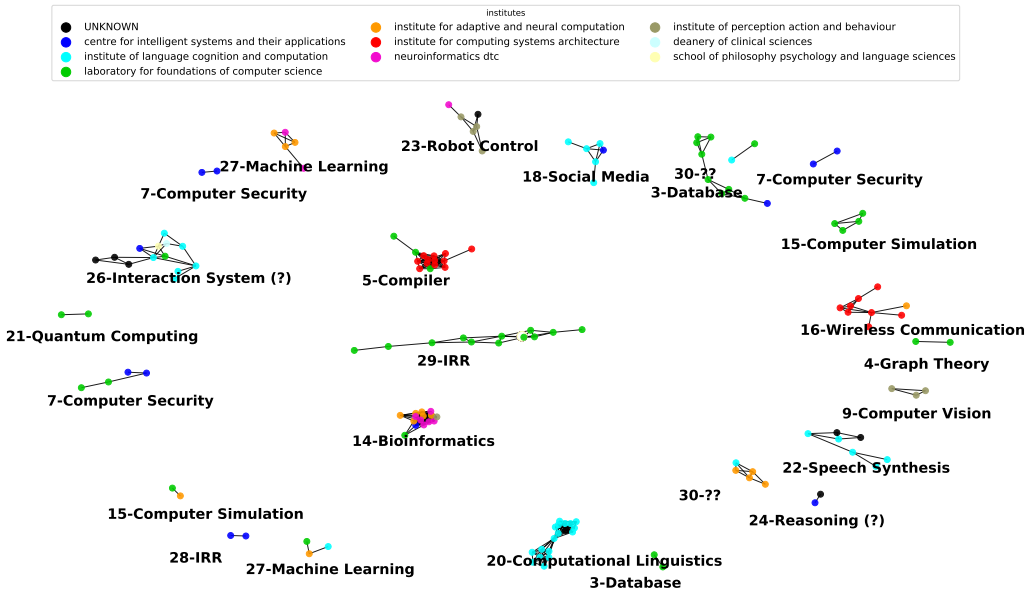


Figure 1.1: A topic-similarity network creates an edge between pair of researchers if the underlying topics discovered from their publications are similar. We retain edges that are significant in, consequently observing topic-based communities where members have publications that reflect common topic of the group. In this figure, we observe `topicnet-6yr`, that is derived using topic model from Informatics publications between 2012 and 2017 (`tm-6yr`). We labelled each of the 30 topics in the topic model and provide each community in the resulting network with a label that best describes the common topics found in their publications. A complete exploration of topics discovered is described in Section 4.3.

We extends our investigation of collaboration patterns by deriving and analysing similarity network created using text data underlying a collaboration network - **topic-similarity network** - providing an attempt to make up for where collaboration network falls short. Simply put, this network ask the question of how closely related the work of two scientists are and accounts for all scientists in the dataset. By representing a scientists' publication as topics of computer science (Blei et al., 2003), and approximating a network similar to our ground-truth collaboration network, we provide an alternate view of collaboration in Informatics (Figure 1.1) .[1]

In the following sections, we describe the hypothesis that motivates this project and related work in each domain in Sections 1.1 and 1.2 respectively. In Section 1.3, we highlight our findings and layout of this report in Section 1.5. A summary of our contributions is listed in Section 1.4.

[1]The term 'topic-similarity network' should not be interpreted as a network with topics as vertices and edges between topics describing the relationship between topics. Such a network is known as *topic network* in Blei and Lafferty (2007) where a weighted edge represents correlation between pair of topics discovered from a collection of text.

1.1 Hypothesis

A collaboration network is derived from bibliometric information of publications but underlying text data is seldom use to contextualise the derived network. Further, publications with an author count of one are ignored. In the case where we are only interested in publications involving subset of scientists from the network, effective publications used is even more scarce. While standard network analysis tools provides a good understanding of our collaboration network, these networks are usually large where number of nodes is in the order of ten thousands. For small networks, we can do better by leveraging on collection of text data underlying a collaboration network. In the following subsections, we gather some intuition on why using text data might work for scientific collaboration networks.

This project stems from the intuition that:

1. Scientific collaboration network is a basis for finding expert groups as participating in collaborative work is frequently due to common interests in some field of study.

2. Experts in a particular field of study would use vocabularies that are dissimilar to those from other fields (expert vocabularies). For example, research groups on computer networks would commonly use terms such as 'routing', 'gateway', and 'redundancy' more than research group focusing on operating systems. This is possible for scientific publications as publications are usually written using terms descriptive of specific field of study.

3. Consequently, if we connect researchers together according to similarity of topics discovered from their published works - a topic-similarity network - then we would observe a alternative representation of a collaboration network. **Such network should preserve properties observed in collaboration networks.**

How do we quantify if two networks are similar? We hypothesise that networks are similar if both obey the principle of homophily - that nodes (researchers) in a network are more likely to be connected to each other if they are from similar background. Here, background refers to an attribute that we can quantify for all the nodes in the network. In the case of the collaboration network, this might be one's institute membership, number of publications, or number of years in the network (age). For example, the collaboration network may exhibit homophily by age, as we observe nodes are connected if they are of the same age more frequently than if connections were to occur at random. However, this may not be the case in topic-similarity network if we observe the converse, where edges between nodes of the same age occur randomly. Consequently, both networks do not agree in homophily by age.

Communities detection and homophily We can further investigate the two networks by applying the homophily test on the same network that is being partitioned by a clustering algorithm .[2] Returning to our afore example before, we remove edges from

[2]Naturally, if we test for homophily based on the community each node is in - then our network

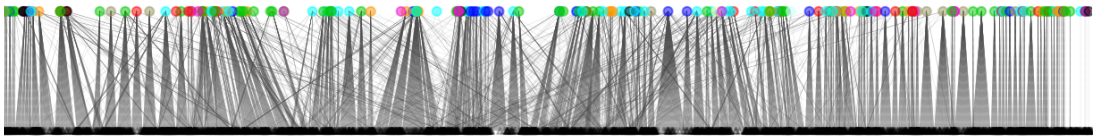


Figure 1.2: Every scientist in the collaboration network affiliates with documents that he/she participated in. Such is an instance of an affiliation network where nodes either represent individuals or documents and can be represented as a bipartition between type of nodes. The black triangles represent documents in our dataset, nodes representing individuals form various institutes in Informatics.

each network until it reaches the partition output by a clustering algorithm. We then carry out the same homophily test using 'age'. With a network where edges connecting communities are removed. If homophily is indeed present, this clustered network should produce similar result on the homophily test.

1.2 Literature review

Two underlying domains in this project are: 1) Network science - a field concerned with actors interacting with each other, quantifying these relationships using graph theory, statistical and sociological methods; 2) Topic modelling - a statistical inferential method used to uncover topics, organise, and summarise collections of text data. Both domains are well-established field with copious amount of research papers published. We provide an overview of theories and methods relevant to our project in this section.

1.2.1 Network Science in Collaboration Network

Interests in collaboration network could possibly be traced back to the establishment and usage of Erdõs Number, to measure the distance between an academic and Hungarian scientist, Paul Erdõs, with the aim to codify the relationship of mathematicians (Goffman, 1969). This also gave a sense of how well-connected researchers in this field are connected to a popular individual in the field. Coupled with sociologists's interests in small-world phenomenon (randomly chosen individuals in the world are six degrees apart in expectation) (Travers and Milgram, 1969; Granovetter, 1973; Watts and Strogatz, 1998; Kleinberg, 2000; Amaral et al., 2000), various models were developed alongside with the aim to characterise other phenomenons such as preferential attachment (Barabsi and Albert, 1999; Jeong et al., 2003), clustering in social networks (Newman, 2001a), community structure and its detection (Girvan and Newman, 2002; Newman, 2016), and scale-free properties (Barabsi and Albert, 1999).

The interest in how scientists from the same field work together is of particular interest too. Improving on the idea of using the Erdõs number, and enabled by the presence of

would certainly obey the principle (since there is no outgoing edge across communities).

online repository of academic papers such as CiteSeerX, arXiv, Google Scholar, and dblp, large scale information retrieval and processing of bilbliometric data is no longer an impossible and daunting task. This made possible network analysis on scientific collaborative work in various academic fields, such as in Computer Science (Franceschet, 2010, 2011), as well as allow for theories to be tested and revised (Tomassini and Luthi, 2007). Overall, collaboration network provides a mean to study real-world networks and put theories to test. It is a well-studied field and remains one that is constantly used to test predictive models.

1.2.2 Networks statistics and distribution

Using essential properties of networks such as vertices and edges, we can quantify networks for comparative analysis and classification. Adopting the nomenclature in Brinkmeier and Schank (2005), we can use global or local variables to compare networks. These variables can be described as single-valued statistics or distributions (range of possible values in the domain) presented as a distribution plot or histogram. The work of Medina et al. (2000) on the WWW led to observation of power-law distributions ($y = kx^{\alpha}$) when comparing number of nodes against degree, eigenvalues of adjacency matrix against rank, and number of pairs against pairwise (reachable) distance (hop plot). Kaiser (2008) evaluated different measurements on cohesion in networks: number of connected components (connectedness); transitivity as a global measure of connectivity in the graph; distribution of clustering coefficient against rank; number of leafs and isolated nodes. To quantify the *importance* of nodes, a range of centrality indices measuring different criterion were proposed, e.g. betweenness centrality and eigenvector centrality where the former can be regarded as extent of control over information flow between nodes while the latter measure linkages between important nodes (PageRank being a variant) (Freeman et al., 1979; Newman, 2008). Using notable characteristics of an empirical networks, we can classify it according to well-known phenomena. For instance, a small world network is present if the network has high clustering coefficient but small diameter (Newman et al., 2002); if given a directed network, we can compare the distribution of vertices in-degree and out-degree to observe the presence of hubs; possibly classifying it as a network with preferential attachment (Barabsi et al., 2002).

1.2.3 Communities in Networks and homophily

Clustering, or community detection, is a process of decomposing a network into natural groups (clusters or community) resulting in *high intra-cluster density and inter-cluster sparsity*. Some common methods uses similarity function (or its 'dual' - distances) to measure the cost of clustering nodes. An algorithm would hence aim to decompose a network into groupings that give the lowest cost of separating nodes that are *similar*. Traditional methods include agglomerative and divisive clustering methods such as single, complete or average linkages; spectral clustering - points are first

projected on the eigenspace of the Laplacian matrix, are k clusters are derived using a partitioning algorithm such as k-means (Ng et al., 2001).[3]

Recent methods include the use edge betweenness measure to remove edges in a divisive clustering algorithm setup (Girvan and Newman, 2002; Newman and Girvan, 2003) and modularity maximisation (Newman, 2006). Edge betweenness, similar to betweenness centrality on vertices, quantify the importance of edges. The authors showed high edge betweenness, corresponding to large number of shortest path running through an edge, providing the best measure; edges between communities have high edge betweenness and removing them yielded distinct clusters. Modularity is a quality measure for graph clustering, measuring the ratio of number of edges inter-cluster to intra-cluster. The partition that derives the maximum modularity hence have relatively good partitioning although it was shown that there exists a resolution limit where small clusters are often "marginalised". Fortunato (2010) presents a thorough review of clustering methods. The research space for clustering is large due to the hardness of the problem.

Homophily It is a natural phenomenon that *birds of the feather flock together* - individuals with similar character, background, race, or upbringing tend to congregate and associate with one another. In community detection, we use the strength or presence of edges to partition our graphs; in network homophily we can use node attributes to test if nodes with similar attributes are linked to each other. Both clustering and homophily are not mutual exclusive, but cases where communities do not show homophily should not be surprising.[4] Newman (2003) proposed various measures of homophily using assortative mixing - such as annotated node attributes (e.g. membership in group(s), race, or age), or degree of nodes (similar to "popularity" of nodes). Node degrees assortative mixing (i.e nodes that are popular usual interact with each other) are found to be common in social networks.

1.2.4 Topic Models

A fundamental problem in natural language processing and machine learning is the representation of documents. An analysis of features used to represent text data ranges from simple, yet effective, model such as bag of words, language models, and models learned using deep neural networks (e.g. word2vec). Embedding documents as a distribution of topics is one such model, and Latent Dirichlet Allocation (LDA), de-

[3] Most algorithms in clustering outputs creates hard membership boundary, a node is either in one cluster or not, ignoring the fuzzy membership space for nodes that are near the boundary which are exposed to influence by other clusters.

[4] At this juncture, it is important to note that there is no universally accepted definition of *community* as it depends on the domain of the network. A community in the telecommunication network may refer to geographical regions but refers to tightly knit social groups in social networks which may represent institutes or organisations in a collaboration network.The weight of edges in a social network is akin to friendship bonds (weights).Bonds amongst members of a group will be strong, but weak between individuals from different group. Granovetter (1973) investigated the *the strength of weak ties* from a sociology perspective.

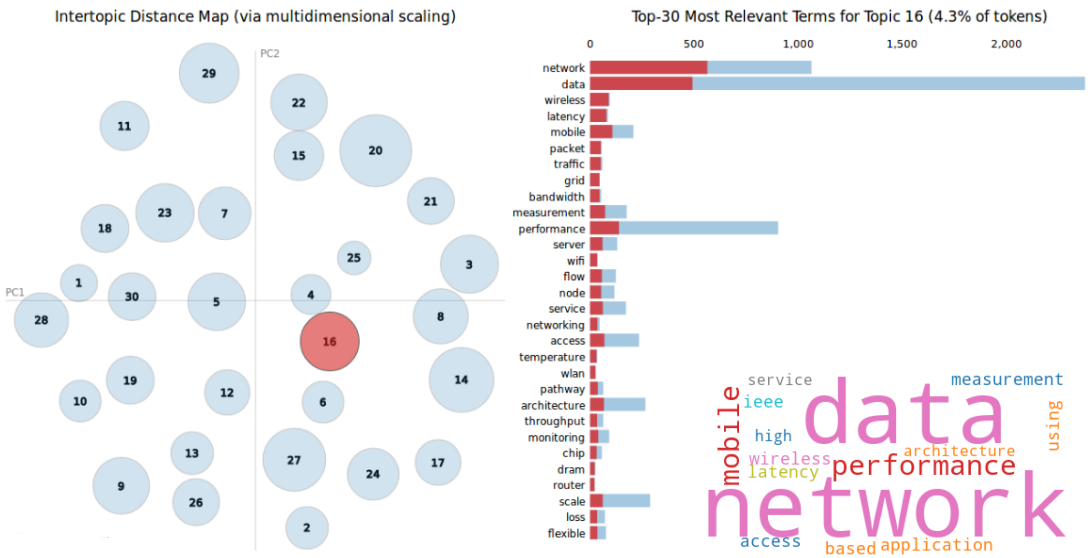


Figure 1.3: Visually, each topic is a distribution of terms that co-occur frequently, where terms that frequently occur are rank higher. We can use a word cloud to visualise the top-15 term where the size of each term is scaled according to its frequency (inset). From the distribution of terms, we can label Topic 16 from `tm-6yr` as *Wireless Communication*. Exploration of topic models derived in this project can be accessed from our companion website.

veloped by Blei et al., is commonly used algorithm that models text probabilistically where topics are characterised by a distribution over words, and documents are a random mixture of topics. The task is hence to discover the posterior distribution of topics and terms.

To approximate the latent topic structure to be close to the true posterior of the collection, multiple topic modelling algorithms have been proposed: mean field variational inference (Blei et al., 2003), Gibbs Sampling method (Griffiths and Steyvers, 2004), collapsed variational inference (Teh et al., 2007), and expectation propagation (Minka and Lafferty, 2012). An online algorithm to handle large collection of documents builds on variational inference is proposed in Hoffman et al. (2010), allowing LDA to be used on large collection in the order of millions. Blei (2012) provides an good overview of probabilistic topic models developed over the years.

1.2.5 Evaluating Topic Models

A topic generated by topic models is probability distribution over the vocabulary and is often hard to interpret as it may simply be a set of words that may or may not correspond to a semantic concept easily understood by humans. Many methods were proposed to quantify semantic interpretability of topics generated by topic models. For instance, Chang et al. (2009) conducted word-intrusion task with volunteers using outputs from topic models [5], and showed that outputs from topic models can be interpreted

[5] In *word intrusion task*, words that do not belong to the set are manually sieved out by participants in experiments. Good topic models will allow participants to point out intrusions with ease (Chang et al.,

by humans and "junk" topics that are incoherent do exists. Newman et al. (2010) proposed automatic coherence measure to rate topics based on word co-occurrence statistics estimated using a reference corpus such as Wikipedia. The underlying intuition is that topic terms that appear within documents suggest semantic relatedness. The similar measure was found to be useful using the same collection of documents the topic model is trained on (Mimno et al., 2011). Aletras and Stevenson (2013) represent topic words as vectors of context features and proposed topic coherence measure of topic words in a semantic space created using Wikipedia. The co-occurrence between two topic word is the cosine similarity between topic word context vectors. They also observed that restricting to the use of topic word space produce the best result. In other words, if many top topic words exists around each other, the cosine similarity will be high and hence the topic is in cohesion. Roder et al. (2015) further evaluate possible way of calculating topic coherence, and proposed a new C_V measurement which they showed is highly correlated to human interpretability of topics.

1.3 Overview of results: Networks in Informatics

Using *Edinburgh Research Explorer*, we retrieved a dataset of publications by researchers in Informatics. From which, we derived three types of networks from our dataset: 1) Bipartite network with two type of nodes - publications and individuals (Figure 1.2); 2) Scientific collaboration network illustrates relationships between researchers (Figure 1.4); 3) Topic-similarity network illustrates the relationship between individuals in the network based on topics derived from each of their publications (Figure 1.1). Our study focused on the latter two networks, with the aim of investigating the research landscape in Informatics by comparing clusters and behaviour of network based on principle of homophily. To the best of our knowledge, this is the first attempt in studying the collaboration network in School of Informatics.

Our study is conclusive that bibliometric data from *Edinburgh Research Explorer* is sufficient to create a robust and accurate collaboration network for School of Informatics. Our code base can be easily extended to retrieve metadata from researchers from other Schools, since the structure and presentation of data is homogeneous between Schools and Colleges in the University. While there exists some limitations for representing researchers who were in Informatics, we find that using a collection corresponding to a shorter time period produces more accurate representation of collaboration network.

We created weighted and simple collaboration networks and find that both are useful to describe collaboration relationships. The weighted network allow us to use use sophisticated network analysis tool where presence of edge weights allow more meaningful detection of properties. For instance, when doing community detection using weighted network, our collaboration network led to higher modularity score on partitioning the network. Further analysis has to be made to investigate if the communities created are

2009)

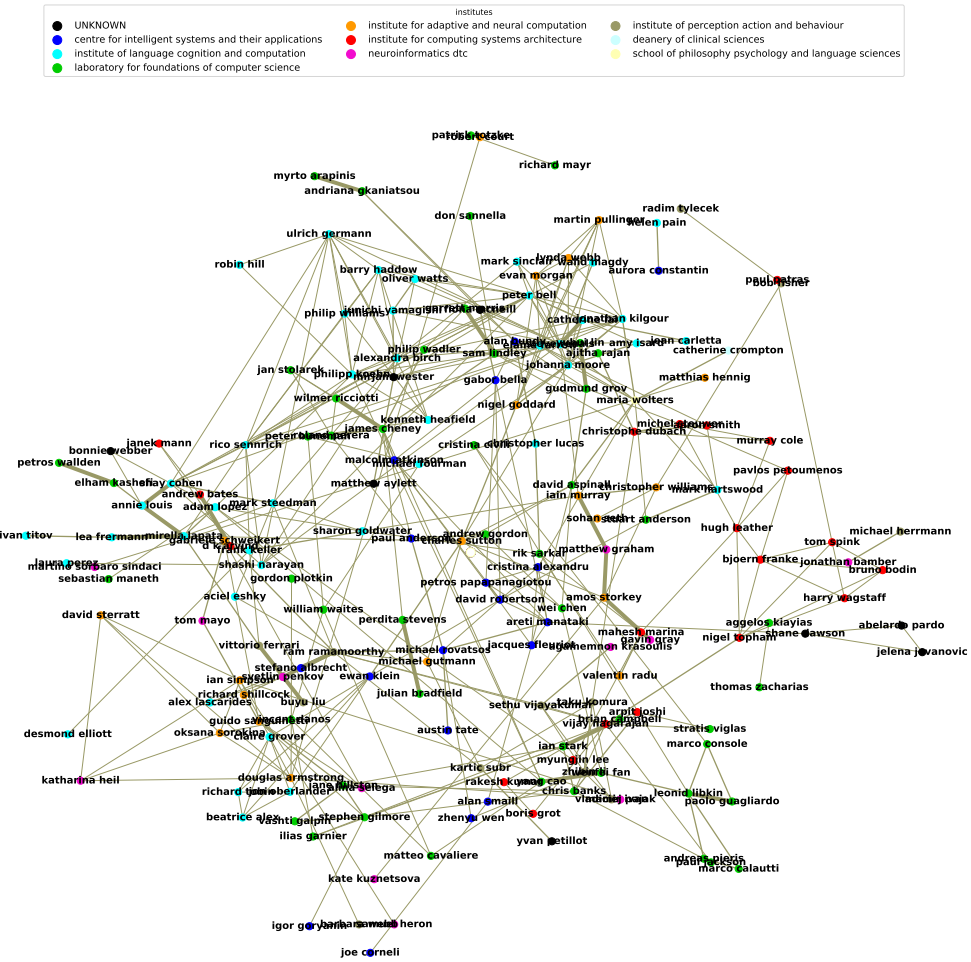


Figure 1.4: `infnet-6yr(w)` where strength of collaboration ties is represented by width of edges - larger edge width represents stronger ties.

meaningful. Aside, our network analysis on Informatics collaboration network is in tandem with other study and we showed that Informatics is a small-world network. In other words, if a researcher is seating in a conference with others from Informatics, a random stranger is at most four degrees away!

Topic modelling using a small corpus of metadata from publications retrieved from *Edinburgh Research Explorer*. resulted in the discovery of about 28 topics that exists in Informatics. We labelled each topic with a semantically relevant topic label and found that they are informative and helpful in understanding research interests across the board (Figure 1.3). We extended the use of topics derived from the collection and proposed a method to qualify relationship between researchers through words. We call this network topic-similarity network and show that it reduced to topic-based communities, similar to applying clustering algorithms on the network (Figure 1.1). This network provides potential for like-minded researchers to find each other in Informatics invigorating collaborations and innovations.

Using both collaboration networks and topic networks, we conducted homophily test

based on institute membership and node degree, where the latter describes the number of collaborators a researcher has. We found that both networks show evidence of homophily by institutes, which means that researchers have a greater tendency to collaborate with one another within institutes. Specifically, we saw strong evidence of homophily in CISA, IANC, and ILCC. We did not observed this relationship for homophily test on node degrees. We found that unlike social networks where gregarious individuals tends to hang out with similar others, this is not the case in collaboration networks. In fact, we saw that researchers with low node degrees tend to collaborate with each other; but when the node degree is high, the opposite is observed. We caution against interpreting this result too deeply, as the power-law nature of degree distribution means that we only observed between 2-3 individuals with high node degree.

1.4 Contributions

Our contributions to the project are as follows:

1. We showed that the bibliometric information on *Edinburgh Research Explorer* is sufficient to create a scientific collaboration network focusing on individuals within the school. It is more robust than those derived from publicly available data, as we can confidently identify individuals from the School using metadata available.

2. Using our dataset, we created scientific collaboration network for researchers in School of Informatics, which to our knowledge is the first instance of it. Our code base can be easily extended to gather similar metadata for other Schools in the University.

3. We modelled the underlying topics/themes in Informatics using Latent Dirichlet Allocation and found about there exists about 28 topics in Informatics.

4. We proposed and implemented an algorithm that constraint the similarity matrix such that topic-similarity network achieves similar size as collaboration network. We show that this can be easily extended to other statistics or a combination of it depending on use case.

5. We compared the derived topic-similarity network with collaboration network, with emphasis on:

    - Communities and network structure (connectedness and degree distribution). The Informatics Collaboration network is a small world network while topic-similarity network is separated into topic-based communities - reflecting localised knit groups expressing similar topics.

    - Network homophily using ground-truth institute membership and degree rank. The result suggests that collaboration occurs within institute more

than across institutes, but inconclusive test results for degree rank due to small sample size.

6. We showed that a topic-similarity network derived using a larger corpus provides an equally meaningful network resistant to biases in dataset.

7. We created an interactive visualisation of networks allow reader to explore the network. This have potential to be extended to allow individuals to explore like-minded individuals in Informatics.

1.5 Outline of Report

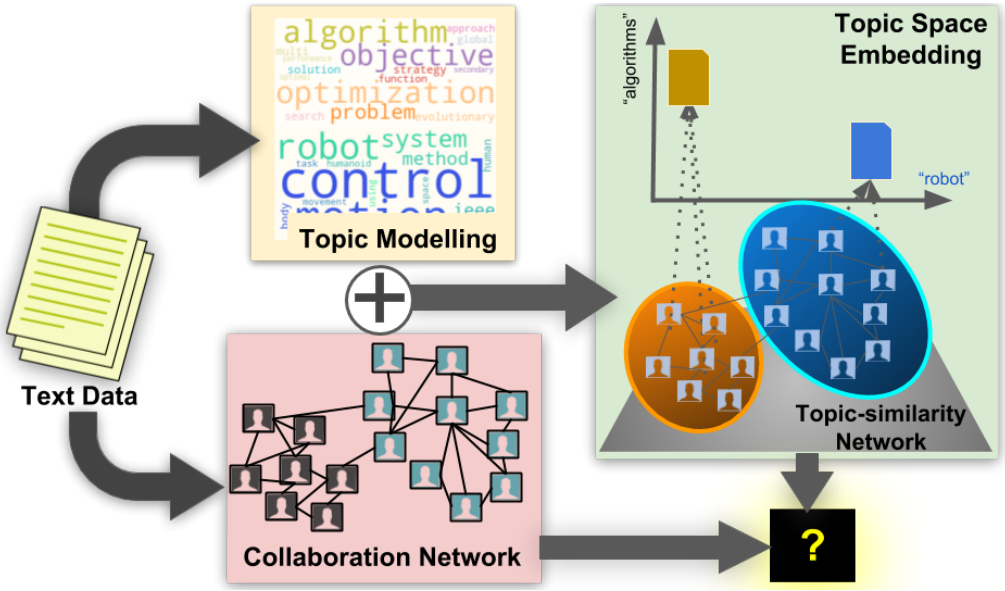


Figure 1.5: This project comprises of five main stages: 1) Gathering and processing of data; 2) Creation of collaboration networks (`infnet`); 3) Generating topic models; 4) Embedding researchers in topic space - creating topic-similarity networks; 5) Comparative analysis of networks.

The structure of this report follows the pipeline of the project (Figure 1.5), starting from the left hand side of the figure.

Chapter 2 **Data Collection and Preprocessing** describes the processes carried out to scrapped *Edinburgh Research Explorer* for bibliometric data for School of Informatics, steps to process the scrapped data, and challenges faced. Descriptive statistics were provided too.

Chapter 3 **Informatics Collaboration Network** describes three variants of collaboration network: `infnet-20yr`, `infnet-6yr`, `infnet-6yr(w)` used in this project. We present fundamental results from network analysis for small-world network and poewr-law distributions.

Chapter 4 **Topics Models** provides an overview of the topic models discovered using LDA - the methodology and experiments conducted to find the best topic model. These topic models - `tm-20yr` and `tm-6yr`- are derived using same collection of documents as the collaboration networks in the previous Chapter.

Chapter 5 **From Words to Networks** describes the process and result of amalgamating a collaboration network and topic model into what we contrive as topic-similarity network.

Chapter 6 **Communities and social influence** evaluates Informatics collaboration network and topic-similarity network using homophily test with respect to institute membership and degree rank.

Chapter 2

Data Collection and Preprocessing

"Garbage in, garbage out!" - an adage from computer science that describes the dependency between quality of results or analysis and data input. In this chapter, we illustrate the process of retrieving data from *Edinburgh Research Explorer* and challenges faced to ensure an accurate depiction of the collaboration networks. We also investigate the distributions of researchers and publications in our dataset.

A previous collection of data on Informatics collaboration was found to be lacking in two ways: an absence of attempts to resolve aliases referring to the same individual and importantly, no metadata of publications was gathered, making infeasible for topic modelling. This motivates the need to gather a fresh set of data. Building on the work done previously, we now describe the process of extending the code base to gather required information from *Edinburgh Research Explorer*.[1]

2.1 Retrieving information from *Edinburgh Research Explorer*

Edinburgh Research Explorer is a one-stop portal for most, if not all, research publications from University of Edinburgh. It contains basic information of researchers from all the schools in the University, as well as metadata of their published work. One could also download published paper from the site, if available. A summary of metadata required for this project is presented in Table 2.1. We used scrapy, a Python framework for web-mining, to create a webscrapper.

Our webscrapper visited the School of Informatics's list of personnel and gathered information of each staff present in the School[2]. Then, it iteratively visits each re-

[1] Our code base use for collecting data from *Edinburgh Research Explorer* is adapted from the code previously used. Modifications have to be made to account for the additional information required.

[2] The list of individual in the school can be accessed from `http://www.research.ed.ac.uk/portal/en/organisations/school-of-informatics(d9a3581f-93a4-4d74-bf29-14c86a1da9f4)`

Metadata	Example
Name	*O'Boyle, Michael*
Position	*Reader*, *Research Assistant*
Organisations	*School of Informatics*, *Institute for Computing Systems Architecture*
URL to personal page	*http://www.research.ed.ac.uk/portal/en/persons/michael-oboyle(1419562d-17ae-4ef2-9014-ca629eed6adb)/publications.html*

(a) Researchers

Metadata	Example
Title	*Four Metrics to Evaluate Heterogeneous Multicores*
Abstract	*Semiconductor device scaling has made single-ISA heterogeneous processors a reality. Heterogeneous processors...*
Date	*Jan 2016*
Collaborators	1. Aliases of collaborators in personal page: *Tomusk, E., Dubach, C. & O'Boyle, M.* 2. Names of collaborators from publication page *Erik Tomusk, Christophe Dubach, Michael O'Boyle*
Publications, Journal, or Conferences	*Machine Learning for Signal Processing* *ACM Transactions on Architecture and Code Optimization* *International Meeting for Autism Research*
URLs	1. Publication *http://www.research.ed.ac.uk/portal/en/publications/four-metrics-to-evaluate-heterogeneous-multicores(4dff484f-21ef-4554-9071-3422481cc41a).html* 2. DOIs *http://dx.doi.org/10.1145/2829950* 3. PDF *http://www.research.ed.ac.uk/portal/files/23696635/-metrics_paper_author_1.pdf*

(b) Publications.

Table 2.1: Metadata required and examples from *Edinburgh Research Explorer*. Each researcher and publications have an unique identifier embedded in the URL. This allow us to identify all staff and publications from Informatics accurately.

searcher's personal page, accumulating information on published publications. Figure 2.1 presents a process flowchart of our webscrapper visiting a member of Informatics's list of publications and a publication's webpage on *Edinburgh Research Explorer*.

scrapy provides customisable pipelines, enabling processing of raw data before storing it in separate files. This automation is especially useful in the following scenarios:

1. **Removal of duplicated publications.** It is common for members from the School to collaborate amongst themselves. Hence, Scrapy would visit the same publication webpage once for every researcher from the school. We use pipelines to ignore similar publications by leveraging on unique identifiers present in the URLs.

`/persons.html`

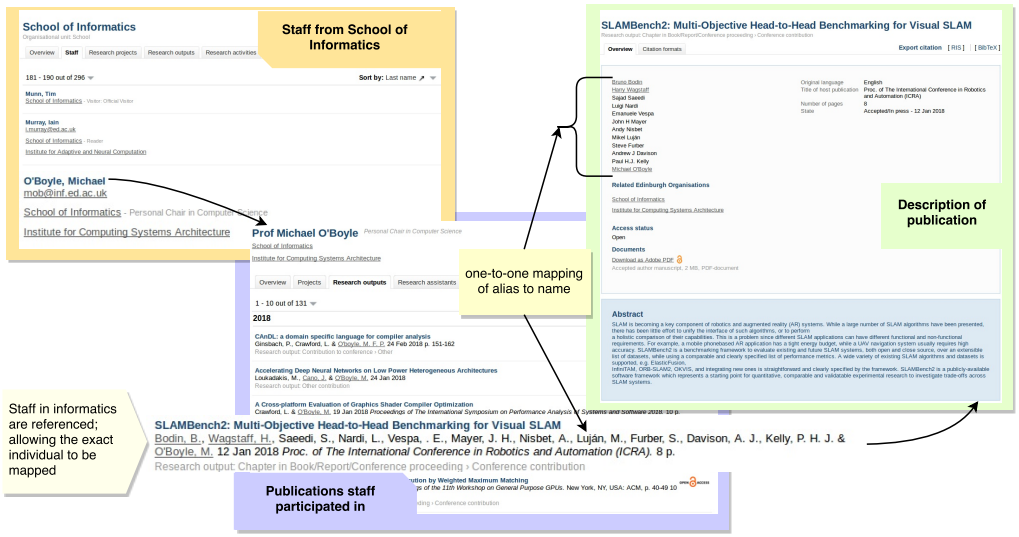


Figure 2.1: From left to right, Scrapy first collects all the URL that links to personal pages of staff from School of Informatics. Next, It visit all the publications that the staff participated in. Before visiting the publication pages, Scrapy collects data about aliases of authors which allows it to create a one-to-one mapping to those in the publication page. We can accurately identify a staff from the School of Informatics by observing hyperlinks embedded in each alias and authors; allowing an accurate representation of collaboration network to be derived.

2. **Standardised representation in memory.** All data are automatically encoded in Unicode-8 to allow international names to be well represented, as well as symbols found in publication titles and abstracts. Case folding was also applied for simplicity.

3. **Gather aliases of individuals.** In his attempt to create scientific citation networks, Newman noted that the aliases of researchers, the contracted form of one's name, complicated the process due to: 1) Different citation styles required for different papers; 2) Different naming convention for authors from different background and culture. Hence, the *last name, first name* convention cannot be strictly enforced; 3) Aliases might change overtime along with changes in one's status, such as addition of middle name.

To that end, our approach is to collect all aliases related to each individual from the portal so that collaboration relationship between individuals can be observed with high accuracy (further elaboration in Section 2.3). To complement the abstracts scrapped for topic modelling, PDFs were retrieved and subsequently converted to text using pdfminer (Yusuke, 2018).

While it is ideal to have perfect dataset with absence of noise, this shows to be challenging. Now, we will visualise the dataset in Section 2.2, shedding light on the incompleteness of the dataset, followed by a discussion on different challenges faced in Section 2.3.

2.2 Descriptive statistics of dataset

Our dataset consists of 288 researchers from School of Informatics and 7,922 researchers outside of Informatics, a collection of bibliometric data from 8,866 published papers between 1969 to 2018. In this project, we focus on publications from 1997 to 2017 (8,028 publications by 212 researchers from Informatics).

2.2.1 Researchers

Amongst the 288 researchers listed on *Edinburgh Research Explorer* we found that 64 of them does not have any published papers on the portal and that they are predominantly 'Visitors' and 'Research associates' of the School.

2.2.1.1 Institutes membership

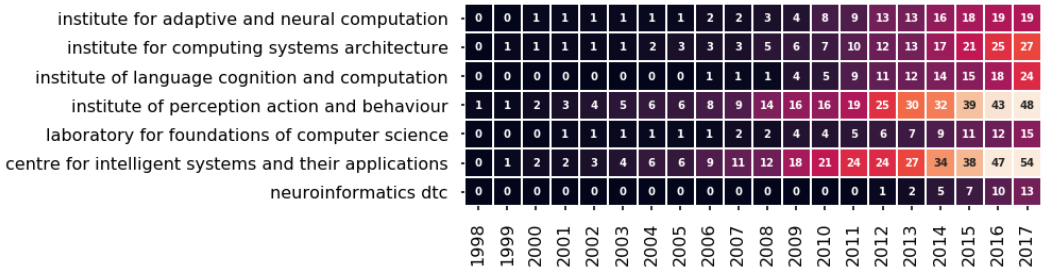
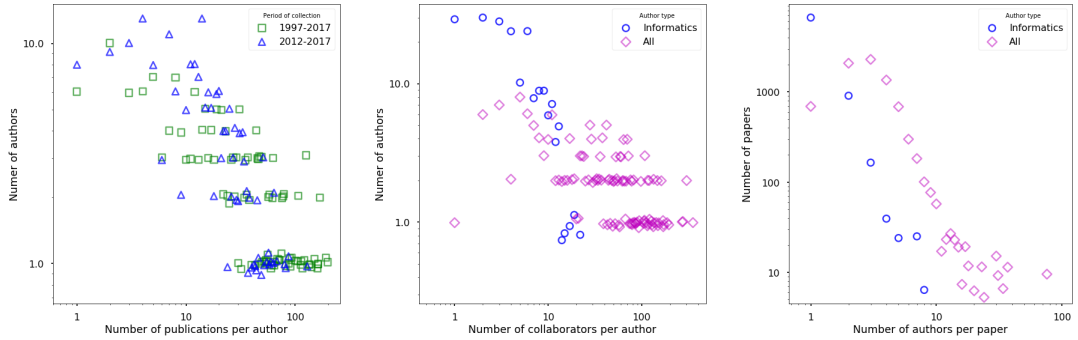


Figure 2.2: Distribution of individuals from various institutes in School of Informatics derived from publications from 1997-2017. Due to incomplete information from *Edinburgh Research Explorer*when the data is accessed, 94.3% of researchers who have at least one publication on the portal have membership information corresponding to one of the seven institutes.

Figure 2.2 illustrates the distribution of membership across the years; each element of the heatmap represents the number of researchers in each institute in the corresponding year. We assign the 'year' a researcher joined based on his/her earliest publication(s) listed on the portal. Although a total of seven institutes made up the School of Informatics, we found that incompleteness on our dataset means that we have a large proportion of researcher does not have any information about institute they belong to. On closer inspection, we found that most of these individuals do not have any publications on the portal between 1997 and 2017. Effectively, for the period of collection we are interested in, there are 212 researchers with 12 not corresponding to any of the seven institute.[3]



(a) Distribution of publication by each author for collection from 1997 to 2017 (green) and 2012 to 2017 (blue).

(b) Histogram of total number of collaborators (magenta) and collaborators from Informatics (blue) for collection from 1997 to 2017.

(c) Histogram of number of authors on each paper. Blue: number of authors from Informatics; Red: total number of authors in each paper.

Figure 2.3: We use log-log plots to compare the distributions where a strong linear relationship between in the distribution. Our collections exhibit power-law relationship for the metrics under investigation, in tandem with findings in Newman (2001b).

2.2.2 Publications

We provide an overview of collaboration patterns in Informatics, using similar methods from Newman (2001b) which showed power-law distributions in 1) Number of publications per author; 2) Number of collaborators per author; 3) Number of authors per paper (Figure 2.3).

1. **Number of publications per author** Fig 2.3a shows that most authors have small number of publications, while few have large number of publications. Similar observations was found for both periods of collection with $\alpha = 2.75$ for collection from 1997 to 2017 and $\alpha = 3.13$ for collection from 2012 to 2017.

2. **Number of collaborators per author** Fig 2.3b shows most authors from Informatics frequently collaborate with more than one other, and have more collaborators outside of the School. If we only count the number of researchers from Informatics per paper, the power-law relationship is clearer - between 1-10 collaborators number of authors decreases gentler compared to more than 10 collaborators. We found $\alpha = 7.38$ for authors from Informatics and $\alpha = 3.89$ overall.

3. **Number of authors per paper** Fig 2.3c confirms our observation and highlights the number of authors per paper. On average, each paper have 3.65 collaborators and 1.14 for collaborators from Informatics, which means in expectation a randomly chosen paper would have about 4 authors where 1 is from Informatics.

[3]Of the 12 researchers, one belongs to 'School of Philosophy, Psychology and Language Science' and 'Deanery of Clinical Science' each, with 10 remaining 'UNKNOWN'.

2.3 Challenges

Additional processing is required to clean up the dataset collected in order to create the collaboration network and for topic modelling. The following two subsections describes the challenges faced while processing the data.

2.3.1 Processing for collaboration network

Ideally, we would like to create a collaboration network representing individuals from the School and external collaborators. This means that we need to accurately identify authors in each publication so that 1) Nodes in our collaboration network corresponds to authors; 2) Edges are created between pair of authors deterministically if they have published a paper. Our fundamental task is hence to develop a methodology that is able to identify authors, which we showed in the previous sections that it is possible to do so by using the unique identifier embedded in the URL in each publication's page. *Edinburgh Research Explorer* however does not have a unique identifier for each researcher outside of Informatics. Our endeavour to create an informative collaboration network hence rest of our ability to resolve a many-to-one problem where possibly multiple aliases points to the same individual.

To resolve this many-to-one problem, we first recognise that there are three sources of information where we can gather data points about aliases corresponding to a researcher:

1. Aliases of researchers in the publication exists in the list of publications when our webscrapper visits the list of publications by a staff.

2. In each publication page, the full name of collaborators can be found which, in most publications, are in the same order as that of the aliases.

3. For researchers from the School, hyperlinks in the list of publications points to their personal page. Which means that the an alias can be directly mapped to the an unique identifier.

Our method is to adopt greedy approach by retrieving all possible data points from these sources of information. Then, a fuzzy match was used to create a many-to-one mapping. This mapping is easy for researchers in Informatics, as we simply have to create a lookup index with his/her unique identifier as key and a set of aliases that points to their unique identifier. It is more tricky for external individuals, and we have to rely on the naming convention between one's name and alias. This means that every time we saw a corresponding name and alias pair, we try to re-create the alias by splitting it into first and last names, and contracting it to citation format. Finally, if the full name is unique, we generate an unique identifier, otherwise we add it to the set of aliases. We found that this approach is limited when the number of collaborators is too many to be listed, resulting in contractions such as '*and 24 individuals*' being used

instead. Also, few names were saw were typeset as *first name, last name* instead which we fail to account in this cases.

Overall, a robust and reliable mapping was derived for researchers in the School, while one that relies on basic naming convention was used for external collaborators. For example, aliases belonging to *Michael O'Boyle* are: "*o'boyle, m. f. p.*", "*o'boyle, m.*", and "*o'boyle, m. f. p. (ed.)*".

2.3.2 Processing for topic modelling

Processing of the PDF and metadata for topic modelling is comparatively straight forward using Natural Language Toolkit (NLTK) (Bird et al., 2009). Overall, the documents are 1) Tokenised; 2) Tokens that belongs to a fixed set of stop words (such as *between*, *into*, *through*, *during*, *before*), special characters, and numbers were removed. 3) WordNet lemmatiser was applied to reduce each token to their dictionary form (e.g. *networking* would be converted to *network*, so will *networked*).[4]

Additional bi-grams and tri-grams are later added to the bag-of-words so that frequent words that collocate are bound together as a token. Such cases were common in description of different areas of computer science such as *machine_learning*, *distributed_system*, and *partial_differential_equation*. Hence, using them to represent the document could better the representation of the document; similar documents bearing these tokens would be discovered by the topic modelling algorithm as they co-occurs and are not as frequent as token (e.g. *system*) itself which can be used in many settings.

2.3.2.1 Comparison between PDF and metadata

The difference in quality of text data retrieved from *Edinburgh Research Explorer* (the metadata), compared to those generated from PDF is apparent. The length of the green bar in Figure 2.4(Right) indicates the presence of a PDF that is successfully downloaded for each year. In total, 4,485 PDFs were scrapped, but only 4,462 were successfully converted to text files (55.6% of the total number of publication under consideration). This indicator is however limited to represent the quality of the PDF. By inspecting the length of tokens representing the metadata scrapped (title, abstracts, conference/journal), the distribution across the years are fairly constant (blue line on Figure 2.4 Left). In comparison, the length for the PDFs scrapped (bearing in mind that not all PDF of publications are available), the distribution varies greatly across the years. This possibly would impact our topic modelling algorithm, because of the unequal representation of papers (some more represented, due to the presence of their PDF, than others). We will refer to this in Chapter 4 - Topic Models.

[4]Initially, Porter Stemmer was used, instead of WordNet Lemmatiser, but the tokens generated were hard to interpret and ambiguity were common when discriminating topics discovered. For instance, *informatics* and *information* will be stemmed to *informat* and *inform* respectively.

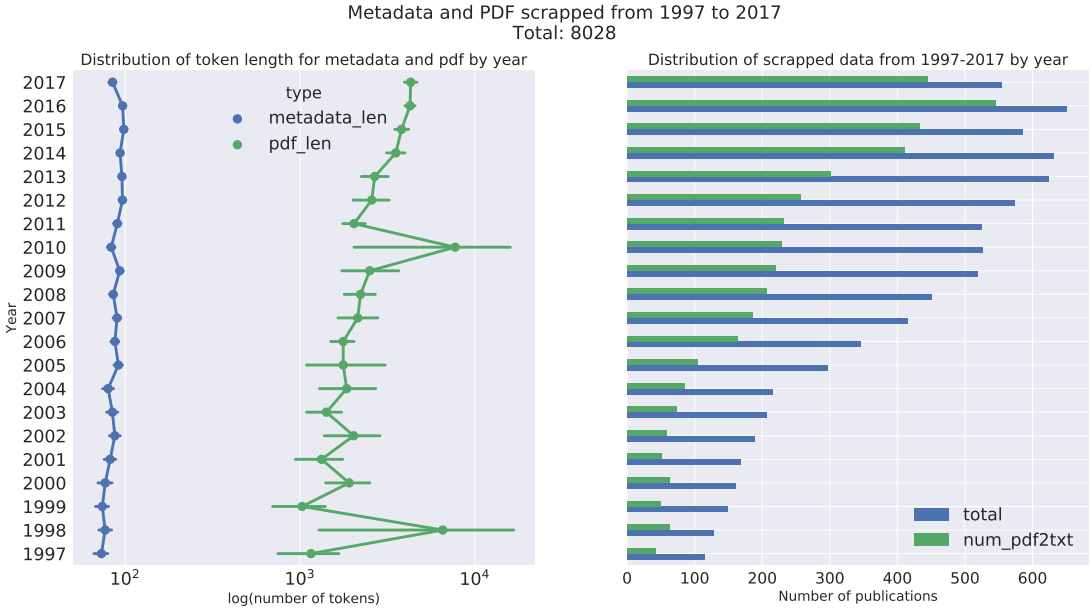
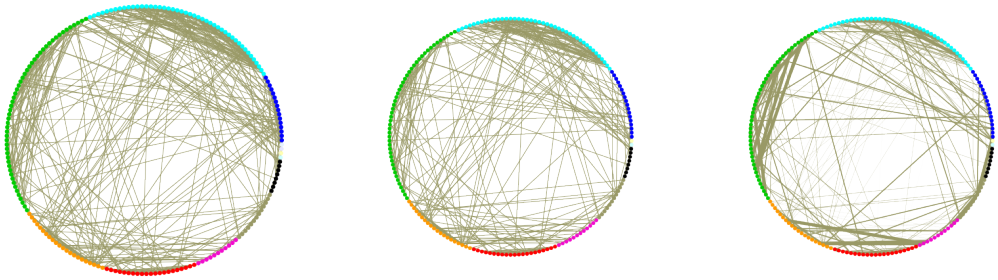


Figure 2.4: ***Left:*** The distribution of token length for the metadata and PDF scrapped provides an indication of the quality of the text data we have. For metadata the distribution does not differ as much compared to that of PDF The differences varied greatly across the year in comparison. ***Right***: Focusing on the published papers from the past 20 years, we see that the number of publications available in the portal increases across the year. The proportion of PDFs available increases too (ratio of green bar compared to blue bar), which could be due to the openness of research publications in the field.

Chapter 3

Informatics Collaboration Network

The Informatics Forum, School of Informatics, houses approximately 250 researchers, and is one of the strongest research centre in the United Kingdom. Divided into seven different institutes, how do individuals collaborate? We seek to answer this question by creating variants of collaboration networks using our dataset: 1) Two simple collaboration network - `infnet-20yr` and `infnet-6yr` - derived from publications from 1997 to 2017 and 2012 to 2017 respectively (Section 3.2); 2) A weighted collaboration network that builds on `infnet-6yr`- `infnet-6yr(w)` (Section 3.3). The excited reader who wish to have an interactive visualisation of these networks should refer to our companion website.



(a) `infnet-20yr` with colours representing institutes in School of Informatics

(b) `infnet-6yr` embedded using the same layout as (a)

(c) `infnet-6yr(w)` is `infnet-6yr` with width of edge representing the weight of edges.

Figure 3.1: Three models of Informatics collaboration network were created and analysed for this project: `infnet-20yr` and `infnet-6yr` are collaboration network that highlight the presence of collaboration relationship between researchers. `infnet-6yr(w)` uses the frequency of collaboration to connotate the strength of collaboration between pairs of researchers. The difference in network size between `infnet-20yr` and `infnet-6yr` is because some researchers who have publications before 2012 but not any after. Since our network is created based on the collection of publications used, such difference is expected.

3.1 Overview

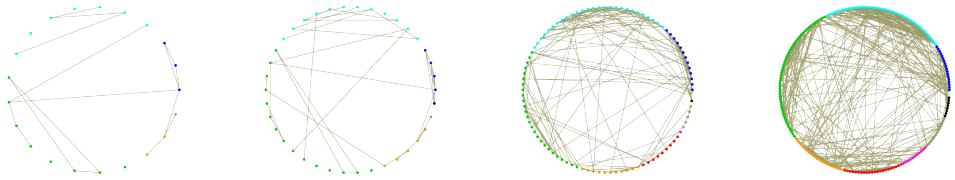


Figure 3.2: From left to right, we can visually observe the increase in size of Informatics collaboration network over time. Each network represent the collaboration network every five years starting from 1997. Each author is given a year based on the first publication we saw in the dataset. Some nodes in the earlier networks does not have any links because the publications that they participated in were before 1997. Overall, we see a network growing in size and collaboration.

A collaboration network have researchers as nodes and edges between pair of researchers are formed if both have collaborated on a publication. From the processed data described in Chapter 2, a total of 8,028 papers were listed in *Edinburgh Research Explorer* from 1997 to 2017, and collaboration network is created using bibliometric data of these publications. Before discussing each model derived, let us understand the limitations of the dataset and the general methodology adopted to create each network:

3.1.1 Assumptions

We assumed that:

- Each listed author contributed equally to the publication since we are more interested in the existence of relationship between individuals than the contribution of each individual to the publications. Hence, the order of appearance of authors in the list does not have any effect on the creation of the network.
- Collaborative relationships between an individuals and the rest of the authors exist and are similar. This implies that given a set of five authors, each individual would have four edges to the other authors, signifying their collaboration for the publication, hence producing a clique for each publication. Although, intuitively this assumption may not hold for publications with large number of individuals, since it is impossible to have 'meaningful' relationship with all (say 10) collaborators, the number of such publication remains a small fraction of the collection (Figure 2.3b).

3.1.2 Methodology

Each variant of `infnet` is derived from the same dataset, but we consider different time periods to suit our modelling assumptions. In general, the collaboration network

	infnet-20yr	infnet-6yr	infnet-6yr(w)
Model Parameters			
Type	Simple graph	Simple graph	Weighted graph
Years	1997 to 2017	2012 to 2017	2012 to 2017
Publications used	1,126	656	
(% total)	(14.0%)	(18.1%)	
Descriptive Statistics			
Nodes	195	184	
Edges	471	361	
Average Degree	4.83	3.92	
Connected Components	4	6	
Average clustering coefficient	0.497	0.603	0.055
Giant Connected Component			
Nodes	189	174	
(% total)	(96.92%)	(94.57%)	
Edges	468	356	
Diameter	10	11	
Average shortest path	4.07	4.99	0.65

Table 3.1: Summary of parameters used to generate networks. `infnet-20yr` present an overview of network using almost all the publications in our dataset, while `infnet-6yr` and `infnet-6yr(w)` uses publications from the past six years (2012-2017) as the quality of the data during this time period is better. For average shortest path in `infnet-6yr(w)`, weights are taken into account when calculating the average shortest path. This does not mean that individuals in the network are 0.65 hops away from each other on average.

is created by iteratively inducing a clique for each publication. Then, the network is created by combining them together. The parameters used and descriptive statistics of each network is summarised in Table 3.1.

3.1.3 Limitations

There are two limitations of our dataset that is pertinent to creation of Informatics network: 1) We are unable to accurately define the presence of external researchers in our network due to the shortfall of information; 2) A network derived over a longer period of time is limited in usefulness, as we are unable to concretely ascertain presence of members of Informatics and the number of publications prior to 2012 is significantly less.

3.1.3.1 Accounting for external researchers

The network derived would only include researchers from School of Informatics, despite the presence of information (names and aliases) of external collaborators. We define a collaborator as external if his/her alias is not present in the list of staff on *Ed-*

inburgh Research Explorer. This exclusion is necessary due to the absence of complete information about external collaborators in the following areas critical to the creation of collaboration network:

1. The publications scrapped are highly likely to be only a fraction of the complete collaboration activities participated by an individual. This means that compared to researchers from informatics, external researchers are only partially represented in our dataset.

2. Name aliasing problem we saw in Section 2.3, is expected to be present but without a concrete solution to ascertain the accuracy of the mapping from aliases to individual.

Hence, discounting the networks of external collaborators allow us to derive an accurate version of the collaboration network emphasising researchers from the School of Informatics. This also means that a significant proportion of the papers with only an author from the School will not be considered in the derivation in the collaboration networks. The weighted collaboration network, `infnet-6yr(w)` provides a means of codifying the presence of external collaborator by taking into account the total number of researchers, internal and external, participating in the paper.

3.1.3.2 Movement of researchers overtime

Another limitation of collaboration networks derived can be attributed to an inherent limitation of using *Edinburgh Research Explorer* as our source of data. The list of individuals present in the school is of paramount importance in determining who we interpret as internal or external individuals. A publication may be participated by members from Informatics who have left the School (such as PhD candidates), resulting in them being considered as external collaborators. This is evident in Figure 3.3, where blue bars shows the volume of publications listed in the portal where none of the authors are from the School. In addition, we notice the number of publications with only one collaborators from Informatics is significant in our dataset (the volume of orange bars in Figure 3.3). These publications will not contribute to our collaboration network because it does not have at least a pair of researchers from Informatics.

Across the years, the number of collaboration increases, as denoted the volume of green, pink and purple bars corresponding to the more than one collaborators from Informatics. The inset in Figure 3.3 highlights number of publications with more than 5 authors. Two observations can be made: 1) Collaboration increases over time in Informatics; 2) The number of publications we retrieve for the past six years is comparably more than prior to 2012.

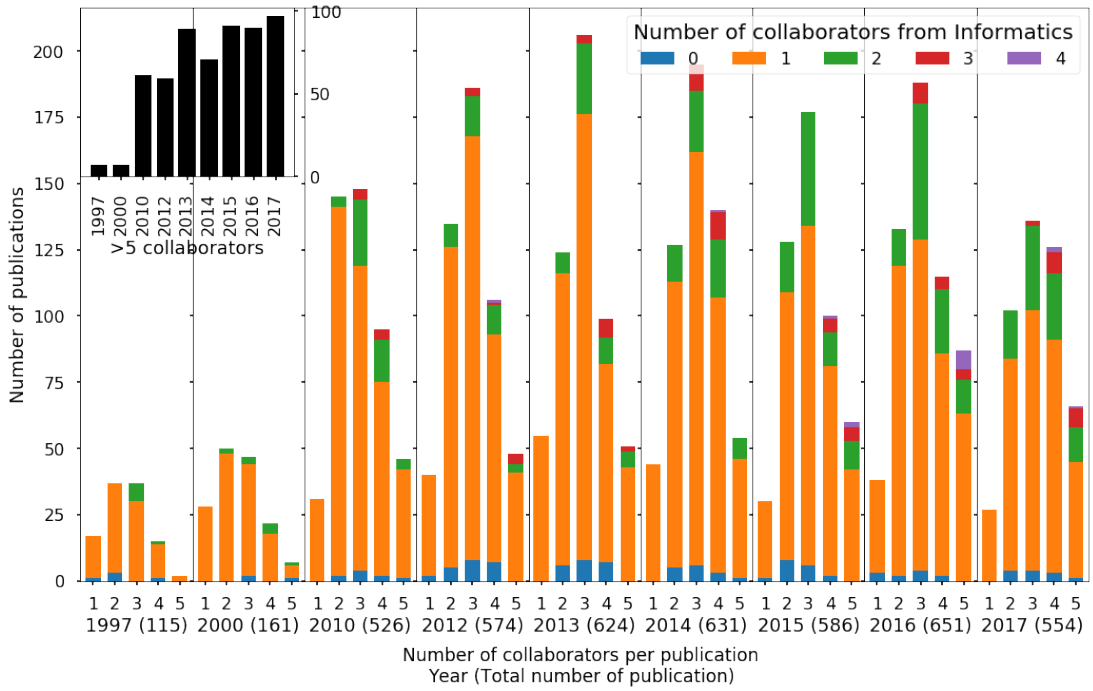


Figure 3.3: In Figure 2.3c, we saw the distribution of publication by number of collaborators which follows a power law - publications with many authors are less common. Here, we sample the distribution of collaborations from 1997, 2010, and 2012 to 2017, and aim to understand the mixture of collaborators from Informatics. Each publication may have one or more authors who may or may not be from the School. The mix of individuals from the School of Informatics (coloured bars) shows this relationship. Inset: Number of publications with more than 5 collaborators, showing the increasing number of publications with large number of collaborators.

3.2 Simple collaboration networks

In this section, we focus on `infnet-20yr` and `infnet-6yr` which describes the collaboration relationship from 1997 to 2017 and 2012 to 2017 respectively. Both networks are simple networks because the relationship between pairs of individual is binary - the presence of collaboration or not, and ignore the existence of multiple edges (multiple collaborations) between each pair of individuals.

3.2.1 Comparison between networks

Comparing `infnet-20yr` and `infnet-6yr`, it is surprising that although the number of publications in the latter is more than half of the former, the number of nodes (195 *vs* 184) and edges (471 *vs* 361) did not differ as much as we would have expected given a difference of 16 years between the two model. This calls into question on whether the collaboration graph employed in our analysis further down the pipeline should be one that covers a longer period. Let us take a step back and look at possible cause of such a difference.

A possible explanation is that the number of collaborations involving individuals from Informatics is better accounted in *Edinburgh Research Explorer* in recent years, as explained above. Looking deeper, we understand that in 2012, the School decided to make it compulsory for all research outputs to be published on *Edinburgh Research Explorer*. This implies that the network with information prior to 2012 is incomplete at best.

On balance, in the interest of the accuracy of models developed further down the pipeline, **the collection of publications from 2012 to 2017 will be used as our baseline model**. We will nevertheless use the entire collection from 1997 to 2017 for comparison the results.

3.2.2 Network analysis

In this section, we conduct network analysis on `infnet`. In specific we are interested to find out if our collaboration network is a small-world network and in agreement with observations by other network scientists.

Power-law on node degree

The node degree corresponds to how many collaborators each researcher have, which would correspond to number of collaborators per author in Figure 2.3b where we saw that there exists a power-law relationship. This means that most researchers work with a small number of collaborators, while very few prolific researchers have large number of collaborators. On closer inspection, we found that these individuals corresponds to researchers in the School who are more senior with position such as ‘reader’, ‘senior lecturer’, and ‘research fellow’.

Small-world network

Using Watts and Strogatz (1998) definition of small-world: the average length of shortest path d between any two individuals grows logarithmically in the number of nodes, i.e. $d = \log n / \log k$ where k is the network’s average node degree. We found that both informatics network have $d = 3.34$ for `infnet-20yr` and $d = 3.81$ for `infnet-6yr`. Comparing to popular notion of *six degrees of separation*, our networks are much smaller, with researchers being about at most four hops away. This show strong evidence that Informatics collaboration network is indeed a small-world network, in tandem with the observations conducted on a larger computer science collaboration network using dblp Franceschet (2010)

3.3 Weighted collaboration network

Building on our baseline model (`infnet-6yr`), we aim to codify the relationship between individuals by giving it a weight based on: 1) Number of publications; 2) Number of collaborators in these publications. Assuming that collaborators interact with each other for the same duration (and hence developing their relationship), then the strength of the relationship will be inversely proportional to the number of possible relationships in the network. In addition, we also codify a compound effect between pair of individuals where if the pair have more collaboration, their relationship strengthens linearly (Newman, 2001c,b).

Mathematically, given a collection of $|D|$ publications, D, the weight between author i and author j is:

$$w_{ij} = \sum_{k=1}^{|D|} \frac{\alpha_i \alpha_j}{f(n_k)} \qquad , \alpha_i^k = \begin{cases} 1, \text{if author i collaborated in paper } k \\ 0, \text{otherwise} \end{cases}$$

There are various choices for $f(n_k)$ to codify how the relationship is spread across the collaborators. $f(n_k) = 1/(n_k - 1)$ is an easy choice as the author have to spend time equally with $n_k - 1$ other authors (Newman, 2001b). We however chose

$$f(n_k) = \frac{1}{\text{size of clique for paper } k} \tag{3.1}$$

$$= \frac{1}{\binom{n_k}{2}} = \frac{2}{n_k(n_k - 1)} \tag{3.2}$$

We rationalise that decreasing the weight by a factor of $2/n_k$ would give a more accurate representation because with more individuals collaborating, time spent between authors will be less than proportionate (Figure 1.4).

Comparing both `infnet-6yr` and `infnet-6yr(w)` in Figure 3.1b and 3.1c respectively, we can see the difference in width of link between pairs with some thicker than others due to the number of collaborations between individuals and some become too thin to be obvious. This might signify that within the collection, we have multiple collaborations between a pair of individuals; where the collaborations are usually between 3-5 individuals. A more elaborated visualisation of the network is laid out in Figure 3.4 where within institute collaboration is laid out using circles, and cross institute collaboration is laid out as shell layouts. We use the width of edges to represent tie strength

3.4 Discussion

In this chapter, we derived Informatics collaboration network using collection of publications from *Edinburgh Research Explorer*. We presented a method to codify strength

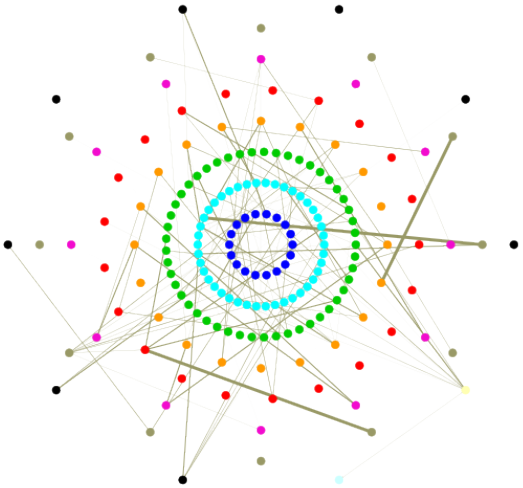
of relationship between researchers by observing the number of collaborators in each paper, and the occurrences of relationship. We can easily extend our implementation to create collaboration networks for other Schools in the University. Our networks exhibit small-world properties, similar to other collaboration networks from other domains. There are many other interesting properties we can investigate in the collaboration network. In Chapter 6 - Communities and social influence, we shall revisit our Informatics collaboration networks and analyse if researchers in Informatics collaborate based on the institutes they belong to (homophily). We also use two clustering algorithm to partition our network, discovering tight knit groups in Informatics.

We also presented limitations of our network, due to the inherent constraint faced from using *Edinburgh Research Explorer*. Further work on Informatics collaboration network can be explored in the following area:

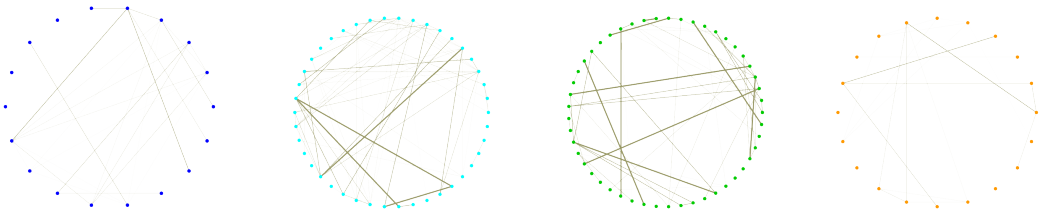
- We presented a series of collaboration network across the year in Figure 3.2. There are much more we can do with this data, such as analysis if preferential attachment is present in the network. This would show if researchers in Informatics would prefer to collaborate with someone who has been in Informatics for a significant period of time. We think that this is possible since when a researcher join Informatics, they would probably be in one of the seven institutes and be under a research group led by a researcher who is more senior.

- We can also analyse resilience in collaboration network, since all collaboration networks consists of a giant connected component *(GCC) which accounts for over 90% of the nodes. By removing nodes iteratively, we can observe the fraction of nodes required to disintegrate the GCC into small disconnected components. A resilient network requires high fraction of nodes to be removed. In other words, we ask the question *how many individuals is required to leave Informatics before we observe satellite group of individuals collaborating together?*

- Our research into collaboration stop short of including external researchers. Interestingly, there are many individuals who are connected to the giant connected component away through an external researcher. We think this could be because the bridge linking a new individual from the main network was a member of the School. Hence, a better model of Informatics Network would include attempts to model these external nodes more accurately.



(a) `infnet-6yr(w)` with collaboration (edges) **within** institutes removed. The additional weight shows the strength of collaboration between individuals across the institutes.



(b) Centre for Intelligent Systems and their Applications (16)

(c) Institute of Language Cognition and Computation (42)

(d) Laboratory for Foundations of Computer Science (49)

(e) Institute for Adaptive and Neural Computation (20)

(f) Institute for Computing Systems Architecture (21)

(g) Neuroinformatics DTC (11)

(h) Institute of Perception Action and Behaviour (12)

Figure 3.4: Compared to `infnet-6yr`, this weighted model highlights significant relationships between pair of staff in the School by increasing or decreasing the weights of relationships based on the number of joint publications and the number of collaborators in each publication.

Chapter 4

Topics Models

In this chapter, we will shift gears and discuss how we used topic modelling algorithm to discover underlying topics in our dataset. We start with an example of how topic modelling automatically categorise textual data to gather some intuition.

Instead of using keyword search alone, for instance, a higher level concept search - a thematic search - may be more useful when dealing with a large collection of documents. We may query about the underlying themes that can be elicited in a collection of documents, or the different themes present in each document. This may not be a challenge if we have a small collection, as given finite resources we might be able to digest, classify, and answer comparative questions between two documents. For example, if we have a collection of publications from two journals: *Journal of Machine Learning Research (JMLR)*, and *Theoretical Computer Science (TCS)*, then finding a paper about current state-of-the-art deep neural network architecture would naturally lead us to browse the collection of *JMRL* as opposed to *TCS*, because we know that "deep neural network" is a theme of "machine learning" which is in turn a theme of "artificial intelligence".

Now, we know these themes and their relationships because we study in the field of computer science and our knowledge about different fields in computer science is aiding our (automatic) abstraction in the little example. If, however, we were given these tasks with a collection of, say, publications from *Digital Humanities* and *Journal of Asian Studies* we may perform badly! An appeal of topic modelling algorithms is they do not require any annotation or label of documents (which often would require a domain expert to do so); statistical structure of documents are illuminated by using the document text itself (Blei, 2012; Blei and Lafferty, 2009).

Formally, topic modelling is an unsupervised machine learning technique that automatically discover 'topics' - a list of terms with high probability of collocation - in a collection of documents. Specifically, we use Latent Dirichlet Allocation (LDA) - a generative probabilistic model - which aims to infer 1) Probability distribution of terms for each topic and, 2) Probability distribution of topics for each document in the

collection (Blei et al., 2003).[1] It is worth noting that all documents in LDA shares the same set of topics and each topic is conditionally independent of the other given the observed documents. While there are other methods that relaxes the independent relationship between topics, such as Correlated Topic Model (CTM) (Blei and Lafferty, 2007; Li and McCallum, 2008)), this project utilises LDA. We discuss the challenges of using other topic modelling algorithm in Section 4.4.

This chapter is organised as such. First, in Section 4.1 a high-level overview of LDA is described. Then, in Section 4.2 methodology adopted to create and choose our topic models will be explained. In Section 4.3, we explore the two topic models derived from publications from 1997 to 2017 (`tm-20yr`) and from a smaller subset of publications from 2012 to 2017 (`tm-6yr`). For comparison, we also created a topic model using bibliometric data of publications available on dblp - an online repository of Computer Science publications. We close the chapter with a discussion of the challenges in Section 4.4.

4.1 Intuition of LDA model

Two assumptions we can make about a collection of documents are 1) Each document is a confluence of multiple ideas; 2) Documents in the collection are heterogeneous and not identical. LDA builds on these assumptions and further assumes that we can describe each idea (topics) with a mixture of terms from a vocabulary. These topics, however, are hidden (latent) and have to be learned from the collection. Our input into Latent Dirichlet Allocation (LDA) are hence the terms present in each document represented using bag-of-words representation over a fixed vocabulary and the number of topics we *suspect* is present in the collection. We can think of the number of topics to be discovered, k, as the granularity of the topics we wish LDA to discover. A larger k would yield more fine-grained topics, and smaller k give rise to larger overarching themes. However, finding the "best" number of topic remains a challenge. Correspondingly, LDA outputs 1) Probability distribution of topics θ_w for document w; 2) Term distribution for each topic β_i for k topics the user wishes to discover.

Blei et al. (2003) shows that it is intractable to calculate the posterior probabilities θ and β; generative model to approximate these distributions given the observed collection is used instead. This means that we approximate the probability distributions and tries to achieve the observed collection by iteratively updating our approximation.

Generative process LDA generates the **terms in each document** in a two-stage process:

Step 1 Randomly choose a distribution over k topics.

Step 2 For each term in the document:

[1] "terms" refers to words that exists in the collection of text, as well as bi-gram and tri-gram that was inferred from the collection. It is commonly known as tokens as well.

a Randomly choose a topic from the distribution over topics in *Step 1*.

b Randomly choose a term from the corresponding distribution over the vocabulary.

The generative process reflects the assumption that each document is a mixture of topics in different proportions (*Step 1*); each term in the document is from one of the topic (*Step 2b*), where the topic is selected based on the per-document distribution over topics (*Step 2a*) (Blei, 2012).

4.2 Methodology - Creating Topic Models

Now, let us understand the derivation of topic models using publications retrieved from *Edinburgh Research Explorer*. In line with our progress in Chapter 3, we aim to generate two topic models subject to same time period constraint used: `tm-20yr` (publications from 1997-2017) and `tm-6yr` (publications from 2012-2017). Additionally, we would also generate a reference topic model using the publications from dblp[2].

The general workflow to generate a topic model is as such:

1. **Input** Collection of documents scrapped from *Edinburgh Research Explorer* is used. We represent each publication as a concatenation of the *Title*, *Abstract*, and the *Publication/Journal/Conference* it is presented in.

2. **Preprocessing** Each document is preprocessed according to Section 2.3.2: 1) Tokenised; 2) Removal of stopwords and non-alpha characters; 3) Lemmatisation using WordNet corpus; 4) Bi-grams and tri-grams generated and concatenated to the list of tokens. The output from this process is a bag-of-words representation of each document.

3. **Filtering** Documents are filtered by year, according to the time period under consideration for each model - either 1997-2017 or 2012-2017.

4. **Building vocabulary** A fixed vocabulary is a set of terms that exist in all the documents present in our filtered collection. The size of the vocabulary is usually in the order of ten thousands, but is usually sparse, as most documents do not possess large proportion of terms in the vocabulary. For storage and computation efficiency, we filter the vocabulary by removing terms that 1) appears in less than 10 documents (rare words), or 2) appears in more than 50% of the documents (common words): The resulting vocabulary is about 2% of the original, and captures terms that are essential but not excessive for each document. This operation can be seen as removing noise from our data so that the LDA generative model pick up essential terms to describe each topic.

[2]dblp if an on-line reference for bibliographic information on major computer science publications, and can be access at `http://dblp.uni-trier.de/`

5. **Topic modelling** We use Gensim (Rehurek and Sojka, 2010) which implements online variational Bayes inference (Hoffman et al., 2010) because it is scalable to large collection of documents. For all the models generated, we used the default hyperparameters, while varying the number of topics to be discovered by LDA.

4.2.1 How many topics?

Before we explore the topic models generated in Section 4.3, let us consider the hyperparameter k - number of topics to be discovered - when creating our topic models. The value of k would have a downstream effect. Mathematically, we can represent each document as a k-dimensional vector instead of a vector of $|V|$-dimension given a vocabulary V. Since k is usually much smaller than $|V|$ ($k << |V|$), k would determine the dimensionality of topic space where we embed our documents in, and the number of parameters required to describe each publication. A large k would led to difficulty in discerning meaningful topics in each document, while a small k would have little discriminative power between documents. Choosing k hence is not an easy task (Arun et al., 2010).

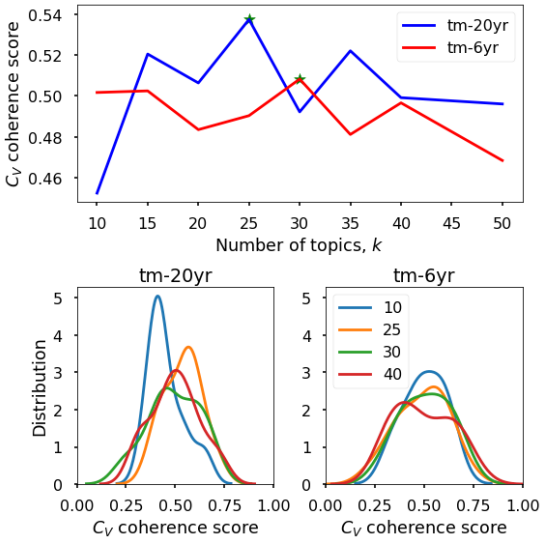


Figure 4.1: We choose the best number of topics based on the highest average C_V coherence scores for `tm-20yr` and `tm-6yr` with $k \in \{10, 15, 20, 25, 30, 35, 40, 50\}$. The optimum number of topic for `tm-20yr` is 25, and 30 for `tm-6yr`. We can also inspect the distribution of topic coherence score for each model, where a higher skewed distribution towards higher coherence score means that the model have topics terms that are coherent. We can observe that both topic model behave differently - `tm-20yr` differs more drastically when k changes compared to `tm-6yr` the latter is more centered around the average C_V score. This means that for `tm-20yr` there are good number of coherent topics and few very incoherent ones; `tm-6yr` have a good spread of topics that are both coherent and incoherent.

Our approach to choosing k relies on the **semantic interpretability of topics** discovered by topic modelling algorithms. In other words, we test if each of the topics generated by our algorithm for some k topics is meaningful and describes a field of study in computer science. In our case, we use the C_v coherence measurement proposed by Roder et al. (2015) to measure the semantic coherence of topics by evaluating each of the top-n terms based on the words it collocates. We evaluate each topic generated using the **top-15** terms and report the C_V coherence score of each topic. The average C_V score guides our choice of k when creating topic models.

4.3 Topic models in Informatics

In this section, we describe topic models derived from our dataset (tm-20yr and tm-6yr) and dblp (tm-dblp) First we present experimental results from finding optimum value of k followed by evaluation of topics in Informatics. We use two

4.3.1 Experimental results for k

We set out to discover the optimum number of topics present for tm-20yr and tm-6yr and observed that tm-20yr is best described by 25 topics, but 30 for tm-6yr (Figure 4.1). Interestingly, although the C_V score for tm-6yr is lower than tm-20yr, we observe that this is caused by the difference is distribution of topic coherence scores in each model. Comparing the distributions, we observe that tm-20yr have larger number of topics that achieve better than its average score, but also have a good number of topics that are performing well below the average. Whereas, for tm-6yr, the distribution are largely centered around the mean, and less perturbed across different k values. This suggests that for the larger collection, the model learned contains more topics that can be considered "junk". We will evaluate this case in the next section.

In comparison to tm-20yr and tm-6yr where the largest collection contains only 8,028 documents, tm-dblp has about 340 times more documents.[3] This presented computation challenges and prevented us from doing a hyperparameter search for the best k (Section 4.4). We estimated that the $k = 100$ would be an ideal value where the topics are not too fine-grained or coarse; and achievable with reasonable compute time (about 36 hours).

4.3.2 Illuminating hidden topics

LDA discovers k topics but it does not provide us with a semantically relevant term to describe each topic. A simplistic way is to use the most probable term in the topic, which is strongly influence by the term count in the collection. Therefore, we instead use the top-15 terms that best describe each topic to create a semantic topic label.[4] There are three ways to present topic-level statistics:

1. In Tables 4.1 and 4.2, we rank each topic discovered according to topic C_V coherence score, and provide the top topic terms that led to our labels. Each term in the list is sorted according to its contribution (posterior probabilities) to the

[3] dblp dataset contains 3,079,007 papers extracted from dblp computer science bibliography (http://dblp.uni-trier.de/). The dataset is extracted by Tang et al. (2008), and contains metadata similar to our dataset from *Edinburgh Research Explorer*. For this project, we use *dblp-Citation-network V10* dataset publicly available here: https://aminer.org/citation

[4] A cautionary note on labelling of topics is that it is extremely subjective to one's experience on the domain(s) present in the document. This means that the topic labels are suggestions and should there be any uncertainty the topic is labelled ??.

topic, where the most common term is on the top left, and the least on the bottom right.

2. Since the distribution of terms per topic is similar to the proportion of occurrences, we can visualise the terms as **word clouds**, scaling the size of each term by the probability of observing it in the topic. Figures 4.2 and 4.3 presents the word clouds observed in each topic model.

3. **LDAvis** provides an interactive way of understanding topic models (Sievert and Shirley, 2014). We invite the reader to interact and explore the topic models generated on our companion website, where three additional properties of topic models is presented (Figures 4.2a and 4.3a presents a static preview of LDAVis).

   (a) Proportion of each topic (shaded circles) in the collection is shown using area. This sheds light on corpus-level statistics.

   (b) Similarity between topics (with respect to the terms they describe) is indicated by the distance between them 2D space. We chose to use multidimensionality scaling so that the distance in k-dimensional topic space is preserved.

   (c) Terms that are most **salient** to the corpus and most **relevant** to each topic. Salient terms are frequent across topics (Chuang et al., 2012), while relevant terms are frequent terms in a topic, but weighted against its frequency in the collection. The latter provides a "local" perspective of the importance of term with respect to the collection [5].

Notation When coming up with topic labels, we find that some topics may not be as coherent as we expect it to be. To that end, we derived two additional topic labels. 1) We use `IRR` to indicate topics that are irrelevant and hence is not a topic of computer science. 2) We use `??` to represent topics that we are uncertain about from the top-15 terms. There are borderline cases where we are able to intuitive come out with a label for the topic but are unsure of because of our lack of experience with these topics in computer science, or there the presence of conflicting terms. When this happens, we use a question mark alongside the suggested topic name, e.g. 'Energy (?)'.

4.3.3 Topics in Informatics

Two models were generated using the same collection of publications but constraint to different time periods - 1997-2017 for `tm-20yr` and 2012-2017 for `tm-6yr`. Table 4.1 and Figure 4.2 illustrates the topic model for `tm-20yr`; Table 4.2 and Figure 4.3 for `tm-6yr`. In general, the topics generated with the best k value found are representative

[5] Relevance of each term w in topic i is a weighted sum of its probability in the topic ($\beta_{w|i}$) and the ratio of the term's appearance in the topic compared to the collection (lift): $\text{rank}(w|i) = \lambda \cdot \log(\beta_{w|i}) + (1-\lambda) \cdot \log\left(\beta_{w|i} / tf_{\text{corpus}}(w)\right)$. Sievert and Shirley (2014) showed that $\lambda = 0.6$ offers the best result and setting $\lambda = 1$ returns the same term ranking observed in the the tables and word cloud.

of topics in computer science, e.g.'Communication Network', 'Graph Theory', and 'Machine Learning'. For both `tm-20yr` and `tm-6yr` we saw two irrelevant topics. In terms of topics that we are uncertain about, there are 2 in the former topic model (out of 25 topics), but 3 in the latter (out of 30 topics). We think that these topics corresponds to meta-topics in computer science.

4.3.3.1 Irrelevant topics (`IRR`)

Both topic models derived exhibit the existence of two irrelevant topics (not a topic of computer science), but are discovered by LDA simply because of copious occurrences in our document collection. Topic 5 of `tm-20yr` and 28 of `tm-6yr` are concerned with the Association for Computing Machinery (ACM)[6] - a computing society headquarters in New York City; Topic 8 of `tm-20yr` and 25 of `tm-6yr` are concerned with publisher Springer[7] headquarters in Berlin/Heidelberg Germany. Both topics exists because of the addition of publication information to represent each document.

Both irrelevant topics have high rank in coherence score, which is not surprising as the co-occurrences of the terms that describe the topics are usually within the neighbourhood of the top topic words. This means that when we observe 'springer', we would also observe 'international', 'berlin' and 'heidelberg' nearby. This should not come as a surprise as the title of a publication usually have the full address of the conference location or publication address. We reproduce two of the publication information listed in *Edinburgh Research Explorer*, that falls into this two 'topic': 'proceedings of the 2016 acm on multimedia conference, new york, ny, usa, acm, acm' and 'artificial neural networks icann 2006, springer-verlag berlin heidelberg'.

4.3.3.2 Does topic modelling illuminates expert vocabulary?

The central goal of topic models for this project is to infer topics of computer science. Each topic, we assumed can be described by a set of words that are specific to the topic. Across the board, the answer is yes - topic modelling does provide an overview of terms that are commonly use in the topic if the topic is interpretable. Here, interpretability refers to the semantic interpretability of terms that co-occur together in a topic.

A limitation to interpreting the topic terms returned by the topic model is that it does not take into account corpus-level statistics of each term. This is the aim of using relevance in LDAvis where top topic terms are ranked with its lift as well. For instance, in Topic 9 of `tm-20yr`, Natural Language Processing, we observe that 'model' is the second most prominent term in the topic. However, when we take into account the expectation of observing 'model', it has a relevance ranking of 12, below terms such as 'sentence' and 'natural_language' which are more useful when describing the vocabulary used the topic.

[6] `https://www.acm.org/about-acm/acm-history`
[7] `http://www.springer.com/gb/about-springer`

We can also answer the question by observing the existence of "junk" topics - topic terms that co-occurs but do not correspond to a topic in the domain. One way to interpret this is through observing the proportion of documents that represent each topic. In LDAvis, this is represented by the area of each topic. In our topic models, we observe that topics that are labelled `??` are usually smaller than others and these corresponds to topics 12, 13, and 25 in `tm-6yr`.

Exceptions to our postulation are present too, as some topics are relatively smaller than others but are coherent. For example, topic 1 of `tm-6yr` rest describes 'graphic_processing_units' and 'gpu' is the second smallest topic, and topic 4 on graph theory is the third smallest topic in the collection. Hence, juxtaposing size of topics and coherence scores, we usually find "junk" topics, but may observe good quality topics.

4.4 Challenges

4.4.1 Using metadata and PDFs

In Chapter 2, we saw that many publications have digital copies (PDFs) freely available online. We intend to leverage on this rich set of text data for topic models under the hypothesis that publications will be better represented, as compared to the simplistic use of the metadata, allowing better and accurate topic model to be generated. However, upon converting the PDF to tokens, we observe that the data is not helpful in two ways:

1. Publications of the early 2000s in the form of PDFs are less frequent than recent publications. This means that our additional tokens will be biased towards more recent publications should we use it for topic modelling.

2. The tokens generated from the PDF using pdfminer (as described in Chapter 2, are inaccurate. This may be due to the presence of tables, algorithms and equations which are frequent in papers from computer science, which confuses the converter to generate tokens approximating the character it saw. For instance we saw a large number of 'sic' which is a representation of a character or word form that the converter fails to recognise. Topic models generated from PDF are also available for exploration on the companion website.

As a result, both topic models (`tm-20yr` and `tm-6yr`) uses only the metadata of publications.

4.4.2 Generating Correlated Topic Model

Correlated Topic Model relaxes the condition that latent topics are independent of each other; pairwise correlation between topics can hence be inferred (Blei and Lafferty,

2007). With this additional correlation information between topics, our model would be stronger, as it can infer related topics. For example, a topic on machine learning would have a stronger correlation coefficient to 'bayesian inference' than say 'computer networking'.

We attempt to generate CTMs so that this additional information can be leveraged for the creation of topic-similarity network. To that end, we experimented with code publicly available by Blei and Lafferty (2007).[8] Our attempts were futile as the machines we were working were not able to interpret the codes few versions away. Consequently, we dropped the idea of generating CTM, and utilised LDA instead.

4.4.3 Large dataset

Working with a large dataset such as dblp, push the machine to the limit when generating LDA models because:

1. Documents has to be read into memory during computation where the dataset is approximately 25 GB in size.

2. Longer time required before estimated posterior probabilities converges.

To that end, we used a multi-core implementation of online LDA which slightly compromise the accuracy of topic model trained but speeds up the process by using most of the core on a machine (Rehurek and Sojka, 2010). The online LDA means that it only requires fixed-size memory and in the end, it took approximately 1.5 days to converge with 100 topics.

4.5 Discussion

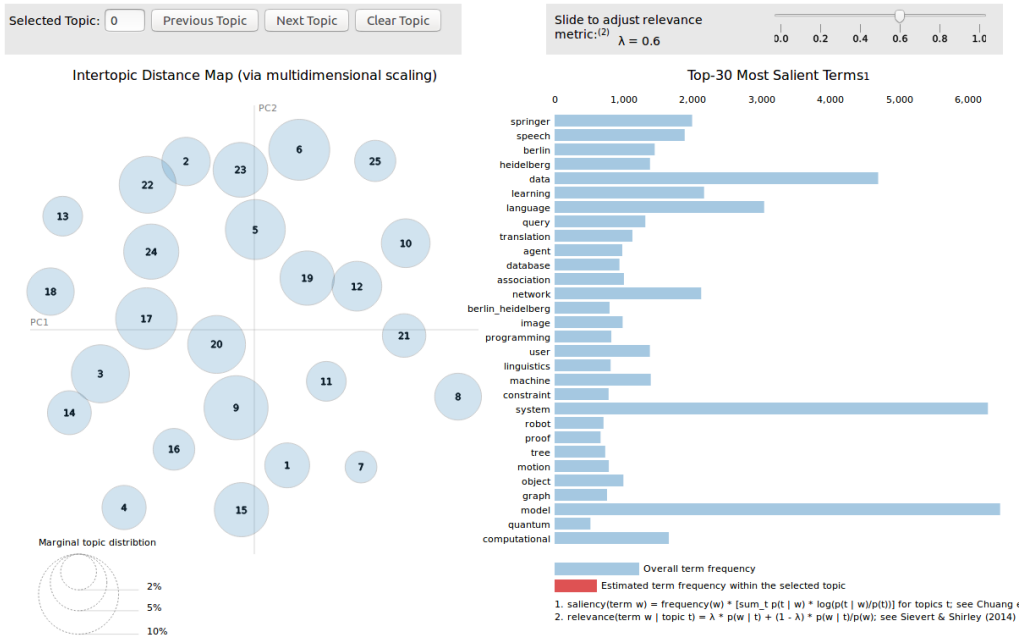
We presented two topic models inferred using LDA which we will use in the next chapter - From Words to Networks to create topic-similarity networks. At this juncture, we note that topic modelling is one of the many available models we can use to create embeddings for documents in a corpus. In comparison with more advanced methods such as using a neural network to learn context vectors, we find LDA appeals to our need as it is easier to interpret.

Future work on topic modelling on documents from Informatics can include analysis on paper distribution between topics - which we assumed is equal for all topics. This would allow us to better model the published papers from Informatics.

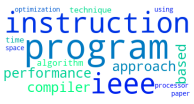
[8] Code for Correlated Topic Model can be accessed at: `https://github.com/blei-lab/ctm-c` .

Topic#	C_V score	Top 15 topic terms					Topic Name
4	0.722	translation system statistical	association machine_translation task	machine association_computational model	computational proceeding association_computational_linguistics	linguistics language computational_linguistics	Machine Translation
8	0.689	springer international springer_verlag	berlin heidelberg_springer paper	heidelberg proceeding gmbh	berlin_heidelberg verlag model	springer_berlin conference system	IRR
15	0.668	model stimulus neuroscience	neural synaptic information	neuron response population	network brain spike	activity cell change	Neuroscience
5	0.622	system application ny	data york ny_usa	acm new_york access	user proceeding computing	new usa distributed	IRR
25	0.611	protein biology using	cell model data	gene network pathway	system molecular biological	expression analysis system_biology	Bioinformatics
1	0.609	object model computer	image conference segmentation	international class computer_vision	publishing method part	springer vision recognition	Computer Vision
3	0.602	speech feature data	model using recognition	system acoustic paper	synthesis based network	speaker voice speech_synthesis	Speech Synethesis
18	0.595	tree automaton given	problem relation regular	xml time transducer	algorithm class grammar	show complexity document	Relational Algebra (?)
21	0.576	language code parallel	programming functional calculus	type system implementation	program programming_language java	semantics level acm	Formal Language
17	0.566	learning machine probabilistic	model bayesian algorithm	data approach neural	method problem parameter	inference machine_learning network	Machine Learning
12	0.566	proof formal property	protocol system reasoning	security theory method	logic verification computer	theorem automated specification	Formal Verification
22	0.556	control ieee humanoid	motion dynamic space	robot based using	system human task	method body movement	Robot Control
14	0.553	data schema source	query answer relational	database acm proceeding	provenance system rule	answering information query_answering	Database System
9	0.551	language natural proceeding	model using conference	corpus dialogue semantic	word natural_language computational	text method grammar	NLP
19	0.528	agent planning based	system approach international	knowledge intelligence paper	ontology reasoning domain	interaction artificial proceeding	Agent Based System
2	0.521	program approach optimization	instruction based processor	ieee algorithm using	performance technique paper	compiler time space	Compiler Optimisation
6	0.504	research design science	project paper challenge	technology system data	web service use	tool computer support	??
23	0.468	network ieee high	performance wireless system	application architecture using	memory cache based	energy mobile show	Communication Network
20	0.447	image based video	data method feature	search model level	task analysis using	visual classification domain	??
24	0.445	model game computer	system modelling rule	process pepa analysis	stochastic state space	algebra time science	Computer simulation
10	0.444	effect eye science	word human two	cognitive processing experiment	model study information	visual reading account	Cognitive Science
13	0.417	distribution bound random	algorithm result number	scheme efficient threshold	function time optimal	signature problem probability	Algorithm
16	0.413	user international child	proceeding information topic	retrieval conference content	system workshop evaluation	social test story	Information Rerieval
7	0.400	constraint consistency formula	logic order complete	dependency model property	temporal problem show	key first xml	Formal Logic (?)
11	0.362	quantum matching state	graph one classical	computation algorithm based	view measurement key	pattern problem protocol	Quantum Computing (?)

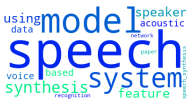
Table 4.1: Topics are ranked according to the topic coherence score where corresponding topics are illustrated in wordclouds in Figure 4.2. We observed two irrelevant topics corresponding to publishers Springer and ACM and two other topics that we are uncertain about. For some topics, such as Topics 7, 11, and 18 we provide a label which we were not completely certain about.



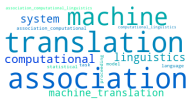
(Topic 1) Computer Vision



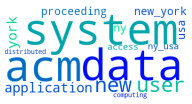
(Topic 2) Compiler Optimisation



(Topic 3) Speech Synthesis

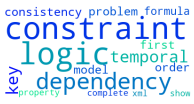


(Topic 4) Machine Translation



(Topic 5) IRR

(Topic 6) ??

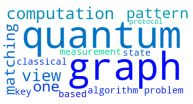


(Topic 7) Formal Logic

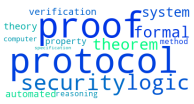
(Topic 8) IRR

(Topic 9) Natural Language Processing

(Topic 10) Cognitive Science -



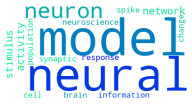
(Topic 11) Quantum Computing



(Topic 12) Formal Verification

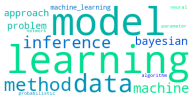
(Topic 13) Algorithm

(Topic 14) Database System

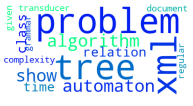


(Topic 15) Neuroscience

(Topic 16) Information Retrieval



(Topic 17) Machine Learning

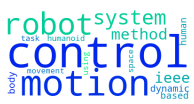


(Topic 18) Relational Algebra

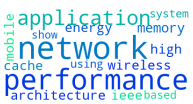
(Topic 19) Agent Based System

(Topic 20) ??

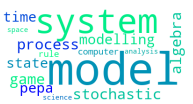
(Topic 21) Formal Language



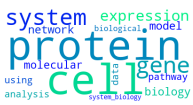
(Topic 22) Robot Control



(Topic 23) Communication Network



(Topic 24) Computer Simulation

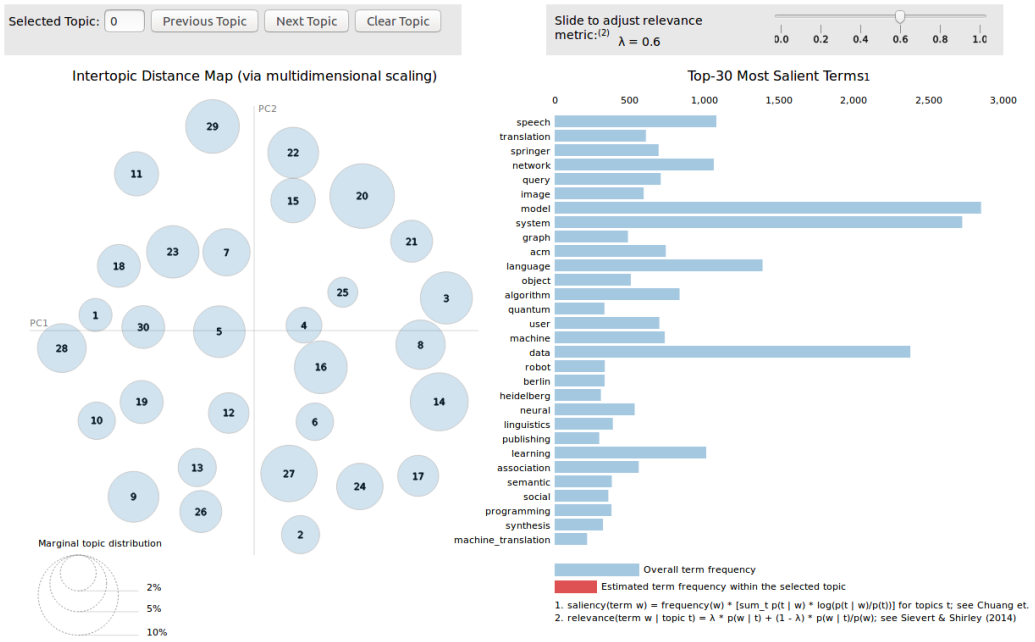


(Topic 25) Bioinformatics

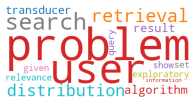
Figure 4.2: Topics and distribution in `tm-20yr`are presented in two different ways: 1) As wordclouds with terms scaled according to the probability of observing it in the topic. We can also view wordclouds as a way of finding terms that co-occurs together in the topic; 2) LDAVis provides an interactive way to explore the corpus and uses relevance to rank terms in the topic.

Topic#	C_V score	Top 15 topic terms					Topic Name
20	0.741	language word text	computational association_computational semantic	association proceeding natural_language	model sentence association_computational_linguistics	linguistics natural based	Computational Linguistics
22	0.687	speech speech_synthesis data	synthesis synthetic based	voice communication speech_communication	speaker paper synthetic_speech	system using text	Speech Synthesis
28	0.686	acm usa conference	proceeding system programming	new ny language	york ny_usa international	new_york agent design	IRR
8	0.662	network speech based	model using recognition	feature deep neural_network	neural ieee data	acoustic training signal	Neural Network
29	0.644	springer conference springer_berlin	international proceeding springer_international	berlin berlin_heidelberg language	heidelberg proof international_publishing	publishing system paper	IRR
9	0.642	image computer model	object method visual	vision computer_vision annotation	class classification training	video segmentation scene	Computer Vision
1	0.636	system gpus implementation	programming parallel unit	level application high_level	high data approach	opencl type accelerator	Parallel Programming
24	0.622	language proceeding representation	reasoning domain workshop	ontology intelligence international	knowledge artificial artificial_intelligence	rule logic semantics	Artificial Intelligence
14	0.594	cell activity population	neuron response analysis	protein brain model	neural gene expression	synaptic information data	Bioinformatics
11	0.577	translation language evaluation	machine paper mt	system statistical shared	machine_translation proceeding data	task english workshop	Machine Translation
17	0.573	algorithm learning value	problem variable degree	time function bound	optimal complexity vector	polynomial probability approximate	Algorithms
18	0.553	social using story	user different conference	web information medium	tweet international arabic	content data topic	Social Media
13	0.546	semantic attribute space	representation category recognition	model domain feature	data visual mapping	learning map alignment	??
5	0.539	system compiler technique	performance software instruction	code based architecture	application acm present	program approach dynamic	Compiler
16	0.526	network wireless measurement	data latency access	performance ieee architecture	mobile using service	application based high	Wireless Communication
27	0.500	model approach machine_learning	method distribution prior	inference data based	learning machine sampling	bayesian probabilistic set	Machine Learning
15	0.489	system analysis simulation	model collective dynamic	modelling approach method	stochastic adaptive formal	process behaviour quantitative	Computer Simulation
23	0.473	robot ieee dynamic	control environment based	task motion learning	provenance humanoid policy	planning system method	Robot Control
10	0.467	memory level fine	program core high	performance processor fine_grained	cache parallel storage	sketch grained design	Computer System
30	0.463	energy design community	research identification science	project practice challenge	social paper approach	technology data future	Energy (?)
21	0.417	quantum category scheme	theory state construction	computation classical measurement	algebra one space	protocol signature structure	Quantum Computing
7	0.415	security application service	workflow cloud attack	protocol computing computation	system data based	privacy process policy	Computer Security
3	0.409	query class show	data answer answering	database problem value	tree graph one	game complexity regular	Database
19	0.393	user interface virtual	motion interactive computer	system environment fish	search interaction based	character information behavior	Human Computer Interaction
26	0.373	dialogue interaction visual	child study participant	people cognitive malware	human behaviour result	learning task design	Interaction System (?)
6	0.369	model using method	parameter approach data	kernel gait input	control dynamic walking	proposed based linear	Modelling (?)
2	0.363	problem algorithm set	user result relevance	search transducer query	retrieval exploratory show	distribution given information	Information Retrieval
4	0.340	graph view based	pattern problem show	data analysis real	algorithm scene graph_pattern	matching using edge	Graph Theory
25	0.285	based error cost	model show chemical	net fusion state	network matter petri_net	method face petri	??
12	0.262	model task speed	data information change	source code using	time result reliability	decision subject annotation	??

Table 4.2: In `tm-6yr`, we observe the same irrelevant topics as in `tm-20yr`. There are more topics that we are uncertain of what it is describing. We think that these topics might corresponds to meta-topics in computer science.

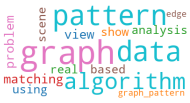


(Topic 1) Parrallel Programming

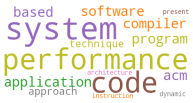


(Topic 2) Information Retrieval

(Topic 3) Database System



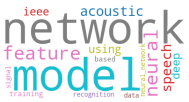
(Topic 4) Graph Theory



(Topic 5) Compiler

(Topic 6) Modelling

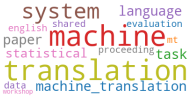
(Topic 7) Computer Security



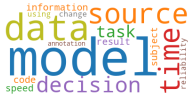
(Topic 8) Neural Network

(Topic 9) Computer Vision

(Topic 10) Computer System

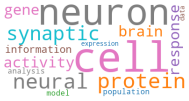


(Topic 11) Machine Translation

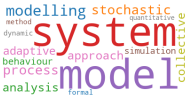


(Topic 12) ??

(Topic 13) ??

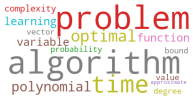


(Topic 14) Bioinformatics



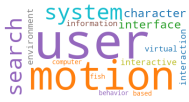
(Topic 15) Computer Simulation

(Topic 16) Wireless Communication



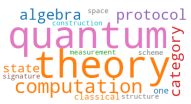
(Topic 17) Algorithms

(Topic 18) Social Media



(Topic 19) Human Computer Interaction

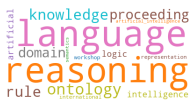
(Topic 20) Computational Linguistics



(Topic 21) Quantum Computing

(Topic 22) Speech Synthesis

(Topic 23) Robot Control



(Topic 24) Artificial Intelligence

(Topic 25) ??

(Topic 26) Interaction Systems

(Topic 27) Machine Learning

(Topic 28) IRR

(Topic 29) IRR

(Topic 30) Energy

Figure 4.3: Topics in `tm-6yr` and distribution of topics across collection of publications from 2012 to 2017. We observe that there are some topics common between the two topic models such as machine translation, machine learning, bioinformatics, algorithms, and computer communications.

Chapter 5

From Words to Networks

A central part of this project is the use of words, from documents, to illuminate the collaborative relationship in School of Informatics. As we alluded to in Chapter 1, a collaboration network derived from publications is akin to social networks. We extend this idea by using topic models to embed researchers in topic space, and create edges between pairs of researchers based on the similarities between their 'research interests'. These 'interests' are represented by topic distributions inferred from collection of publications that a researcher participated in. This similarity network, which we call topic-similarity network, highlights the relationship between researchers using similarity score between topics inferred by a topic model.

So far, we explored collaboration network based on publications from two different periods - 1997 to 2017 (`infnet-20yr`) and 2012 to 2017 (`infnet-6yr`) - and created a weighted version for the latter (`infnet-6yr(w)`). In addition, we used metadata from publications to generate two topic models - `tm-20yr` and `tm-6yr`, each corresponds to a period of publication used to generate the collaboration network. For reference, we used a collection of publications form dblp to derive a topic model too (`tm-dblp`). Since topic model is a critical ingredient for topic-similarity networks, we refrain from using topic models that were derived with tokens from PDF, as noise and bias in the data would introduce additional variations when generating the networks. The combinations of collaboration network and topic model used to derive topic-similarity networks is summarised in Table 5.1.

The layout of the chapter is as follows. In Section 5.1, we describe the steps taken to create topic-similarity networks. In Section 5.2, we illustrate `topicnet-6yr` as an example of topic-similarity networks and provide descriptive statistics of each network, compared to the collaboration network each builds on. Last, in Section 5.3 we discuss implementation considerations and future areas of exploration.

		Collaboration Networks			Topic Models		
		infnet-20yr	infnet-6yr	infnet-6yr(w)	tm-20yr	tm-6yr	tm-dblp
Configuration of networks/models:							
Period	1997-2017	✓			✓		✓
	2012-2017		✓	✓		✓	
Pub	*ERB*	✓	✓	✓	✓	✓	
	dblp						✓
Network and model used for topic-similarity network:							
topicnet-20yr		•			•		
topicnetref-20yr		•					•
topicnet-6yr			•			•	
topicnetref-6yr			•				•
topicnetref-6yr(w)				•			•

Table 5.1: **Top**: We reproduce the parameters used to create each collaboration networks and topic networks for clarity. **Bottom**: topic-similarity network is a derivative of a collaboration network and a topic network, table shows the combinations under consideration for this project.

5.1 Methodology

This section describes the steps required to create a topic-similarity networks as laid out in Algorithm 1. Given a collaboration network $G(U, E^g)$, a learned topic model θ with k topics, and a bipartite mapping $G(U, D, E^{mapping})$ between researchers U and documents D, we create a topic-similarity network $G(U, E^t)$ with researchers as nodes, and an edge e^t_{ij} exists between researchers i and j if the similarity, f, between topic distribution inferred by the topic model θ is greater than threshold ε. We will step through each phase of algorithm 1 in the following subsections.

5.1.1 Inferring researcher's topics

The inference process by a topic model, $\theta(d_i; k)$, estimates the probability distribution of topics given a bag-of-words d_i, representing all the documents of researcher i. To gather the publications a researcher participated in, we use bipartite mapping of documents and researchers $G(U, D, E^{mapping})$ where each individual is mapped to the documents that have co-authored.

Next, we use a learned topic model to infer the probability of each topic observable in this bag-of-words. The output, $\theta(d_i; k)$, would be a vector of length k. This is similar to conducting dimensionality reduction on the document. Since LDA topic model tends to describe every single term it sees, and some tokens are actually common words and have left discriminative power to each topic, we set a minimum probability on the topic probabilities returned. We use 0.01 for `tm-20yr` and `tm-6yr`, and 0.001 for `tm-dblp`. We think that these values are reasonable since the number of topics, k, is 25 or 30 which topic probability are in expectation of 0.04, 0.03, or 0.01 respectively.

Algorithm 1 Using collaboration network $G(U, E^g)$, bipartite mapping $G(U, D, E^{mapping})$ between researchers U and documents D, and learned topic model θ, the topic-similarity network is created in three sequential steps: 1) Inferring topic probability distribution for each researcher; 2) Calculating similarity of topic distribution - equivalent to edges of network; 3) Removing edges so that network is approximately the size of collaboration network.

procedure GENERATETOPICNETWORK($G(U, E^g)$, θ, $G(U, D, E^{mapping})$)
Embed researchers U in k-dimension topic space
  $T \leftarrow$ Matrix of size $|U| \times |k|$
  for all researcher $u \in U$ **do**
    $d_u \leftarrow G(u, D, E^{mapping})$ ▷ Documents matching researcher u
    $T_i \leftarrow \theta(d_u; k)$ ▷ Infer topic distribution (Sec 5.1.1)
  end for
Calculate similarity between researchers
  $E^t \leftarrow$ Matrix of size $|U| \times |U|$
  for all $i, j \in U$ where $i \neq j$ **do**
    $e^t_{ij} \leftarrow \text{CosineSimilarity}(T_i, T_j)$ ▷ Sec 5.1.2
  end for
Prune edges s.t. $|E^t| \rightarrow |E^g|$
  $\varepsilon^* \leftarrow \text{FindBestThreshold}E^g, E^t, 0.001)$ ▷ Refer to Sec 5.1.3
  $E^t \leftarrow \text{SetEdges}(E_t, \varepsilon^*)$ ▷ Eq 5.2
  return $G(U, E^t)$ ▷ topic-similarity network $G(U, E^t)$
end procedure

5.1.2 Calculating similarity f

The similarity between distribution of topics for each researcher is the metric used to create the edges in our topic-similarity network. Cosine similarity is used to compare the topic distribution of each researcher i, j:

$$\text{Let: } \theta(d_i; k) = \vec{a}, \quad \theta(d_j; k) = \vec{b}$$
$$f(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{||\vec{a}||_2 \cdot ||\vec{b}||_2} \qquad f \in [0, 1] \tag{5.1}$$

Here, $f \in [0, 1]$, and is a usual metric used in ranking information retrieval too. If two researchers have the exact same set of publications, then we would expect $f(\vec{a}, \vec{b}) = 1$. The case of $f = 0$ is rare as the learned topic models are noisy and generates probabilities for high level topics that are common to most publications. f's lower bound is 0, instead of -1, as $\theta(d_i; k)$ is a vector of probabilities. This means that the largest angle between any two vectors is 90°. The output from computing the similarity between researchers is a symmetric matrix with elements representing similarity between each possible pair of researcher.

5.1.3 Determining threshold ε

The next step requires us to determine the level at which the similarity between researchers should be considered *significant*. Intuitively, as $f \to 0$, the similarity between topic distribution for a pair of researchers becomes negligible. Since we wish for the topic-similarity network to capture this relationship using the presence of an edge (and weights for weighted network), we simple apply a threshold such that elements in the similarity matrix less than or equal to the threshold is set to 0, denoting the absence of edge between researchers. Now, we explain the choice of threshold $\varepsilon \in [0,1]$.

Algorithm 2 Iterative algorithm to find the optimum threshold ε, such that the number of edges in the similarity matrix E^t approximates that of the ground-truth collaboration network $G(U, E^g)$. ε is increased by δ every iteration.

```
1:  procedure FINDBESTTHRESHOLD(E^g, E^t, δ)
2:      w_min ← smallest value in E^g that is > 0
3:      N_truth ← CountEdges(E^g, w_min)
4:      ε ← 0
5:      n_edges ← TargetNumEdges
6:      while ε < 1 and n_edges ≥ N_truth do
7:          ε ← ε + δ
8:          n_edges ← CountEdges(E^t, ε)
9:      end while
10:     return ε
11: end procedure

12: procedure COUNTEDGES(AdjMat M, threshold ε)
13:     NumEdges ← 0
14:     for all Elements in M ≥ ε do
15:         NumEdges ← NumEdges + 1
16:     end for
17:     return NumEdges
18: end procedure
```

Approximation to ground-truth collaboration network: We use the global network statistics of the ground-truth collaboration network to decide the value of ε and set N_{truth} as the *total number of edges* in the collaboration network **A** so that our topic-similarity network **B** will be of similar size as our ground-truth network. To that end, we used an iterative algorithm that finds the optimum threshold ε^* by iteratively *increasing* ε from 0 to 1 (Algorithm 2). The upper bound of ε is suitable for networks that are weighted too since the resulting weights used in topic-similarity network will be cosine similarity which is bounded by $[0,1]$ (Section 5.1.2).

SetEdges: After finding the the optimum threshold ε^*, each edge e_{ij} is assigned a value by:

$$e_{ij} = \begin{cases} 1, & \text{if } f(\vec{a},\vec{b}) > \varepsilon^* \text{ and } \textbf{binary} \text{ edge} \\ f(\vec{a},\vec{b}), & \text{if } f(\vec{a},\vec{b}) > \varepsilon^* \text{ and } \textbf{weighted} \text{ edge} \\ 0, & \text{otherwise} \end{cases} \tag{5.2}$$

5.1.3.1 Jaccard distance as objective function

Using the number of edges in the collaboration network need not be the only objective function we wish to minimise. One thought is to consider the average jaccard distance between two adjacency matrices from the ground-truth network E^g and topic-similarity network E^t. Since we have two sets of edges that a node is connected to, jaccard distance tells us how different each sets are. This means we are finding the optimum threshold such that the connectivity between edges is preserved, and is the best when nodes are connected the same way as our collaboration network.

Mathematically, let e_i^g represents the set of neighbours of node i in the ground-truth network; e_i^t represents the set of neighbours of node i in the topic-network. Then,

$$J_{dist}(e_i^g, e_i^t) = 1 - \frac{|e_i^g \cap e_i^t|}{|e_i^g \cup e_i^t|} \tag{5.3}$$

$$\overline{J}_{dist}(E^g, E^t) = \frac{1}{|U|} \sum_{i=1}^{|U|} J_{dist}(e_i^g, e_i^t) \tag{5.4}$$

The jaccard distance is 0 when two sets are identical, representing that the collaboration relationships i has is identical to our topic-similarity network. Hence, when the average jaccard distance is low (approaches 0), E^t, our topic-similarity network is most similar to our collaboration network. In our experiments, we also measure average jaccard distance at each value of threshold.

5.2 Topic-similarity Network

We illustrate topic-similarity networks derived in this section. In the interest of space, we provide descriptive statistics of each network in Table 5.2 and elaborate `topicnet-6yr` in full. Next chapter Communities and social influence, we compare compare the topic-similarity network and its corresponding collaboration network.

Compared with the ground-truth collaboration network, topic-similarity networks have more connected components, although smaller in size each are denser. In fact, we postulate that deriving topic-similarity network is similar to using a community detection - where edges connecting between like-minded researchers are preserved, while those

Parameters	infnet-20yr	topicnet-20yr	topicnetref-20yr
Epsilon, ε	-	0.837	0.850
Edges	471	473	474
Average Degree	4.83	5.60	7.35
Connected Components	4	18	18
Average clustering coefficient	0.497	0.687	0.797
Transitivity	0.307	0.756	0.750
Number of leafs	32	19	30

(a) topicnet-20yr and topicnetref-20yr are networks derived by embedding infnet-20yr in topic space of tm-20yr and tm-dblp.

Parameters	infnet-6yr	topicnet-6yr	topicnetref-6yr
Epsilon, ε	-	0.826	0.832
Edges	361	361	361
Average Degree	3.92	4.75	6.07
Connected Components	6	25	18
Average clustering coefficient	0.603	0.785	0.785
Transitivity	0.397	0.826	0.774
Number of leafs	38	39	32

(b) topicnet-6yr and topicnetref-6yr are networks derived by embedding infnet-6yr in topic space of tm-6yr and tm-dblp.

Table 5.2: Summary of descriptive statistics for each of the network derived by embedding collaboration network in different topic spaces. Net for topicnetref-6yr(w) is the same as topicnetref-6yr with the exception of clustering coefficient = 0.700 and the resulting network bearing weighted edges.

that are less similar are removed. **The resulting topic-similarity network is a network with topic-based communities.**

One of the use case of having a topic-similarity network is the ability to map topic labels developed in Chapter 4 to each community, providing a label that represents the cluster. To do so, we calculated the average probabilities for each topic across the cluster, and labelled it using the most probable topic. Figure 5.1 illustrate topicnet-6yr and its label.

Observing the resulting labelled network, some communities are labelled IRR and ?? which means that the most probable topic found is irrelevant or we do not know the label. Since we are only using metadata from the publications, we think that some of the communities may be labelled IRR if the publication's abstract is absent, which means that most of the publications that represents an individual would be related to publisher's information.

If we represent each community found based on the institute where most of its members are from, we would observe that some topics correlates to what we would think the institutes are interested in. For instance, the community interested on Computational Linguistics and Speech Synthesis have most members from *Institute of Language Cognition and Computation*; Wireless Communication and Compiler have most members *Institute of Computer Systems Architecture*; Bioinformatics mainly represented by *Neuroinformatics DTC*; Computer Vision and Robot Control by *Institute*

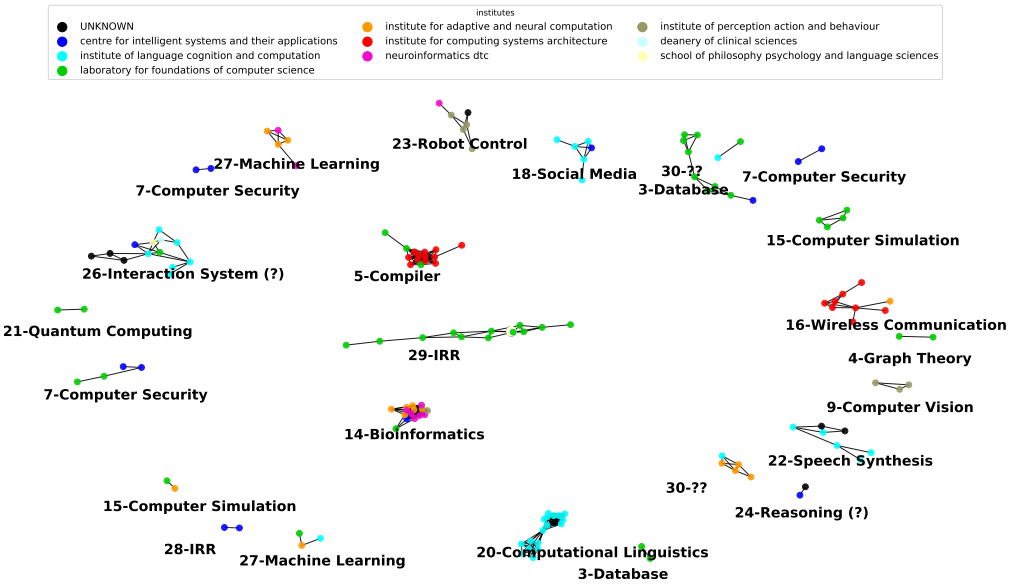


Figure 5.1: We calculate the most probable topic that is being described by each cluster by taking the average of each topic probabilities output by each node in the cluster.

of Perception Action and Behaviour; *laboratory of foundations of computer science* for topics related to Database, Quantum Computing, Computer Simulation, and Graph Theory.

These communities are created based on the similarity between topic distribution, and by observing its relationship with the ground-truth institute, topic-similarity network alludes that institutes may be related to collaborative relationship. We set out to test this in Chapter 6.

5.2.1 Using dblp as a reference collection

In place of topic model trained using the collection of publication from *Edinburgh Research Explorer*, we created topic-similarity network using topic models from dblp. Figure 5.2 illustrates `topicnetref-6yr` with topic labels. Some general observation includes: 1) Some topics derived are different from those we saw previously, which is within expectations since publications used to trained topic models are different, and instead of 25 or 30 topics the model discovered 100 topics instead; 2) The derived networks are generally smaller - which means that there are many isolated nodes. The clustering coefficient remains fairly high, compared to collaboration network, showing that the clusters are relatively dense; 3) Comparing the two topic-similarity networks, we observe that both have similar cluster of Bioinformatics. We also see semantic related topics with respect to what we would observe in an institute. (e.g. IOT and Parallel systems in *Institute of Computing Systems Architecture*).

Since dblp is a much larger corpus that our dataset from *Edinburgh Research Explorer*,

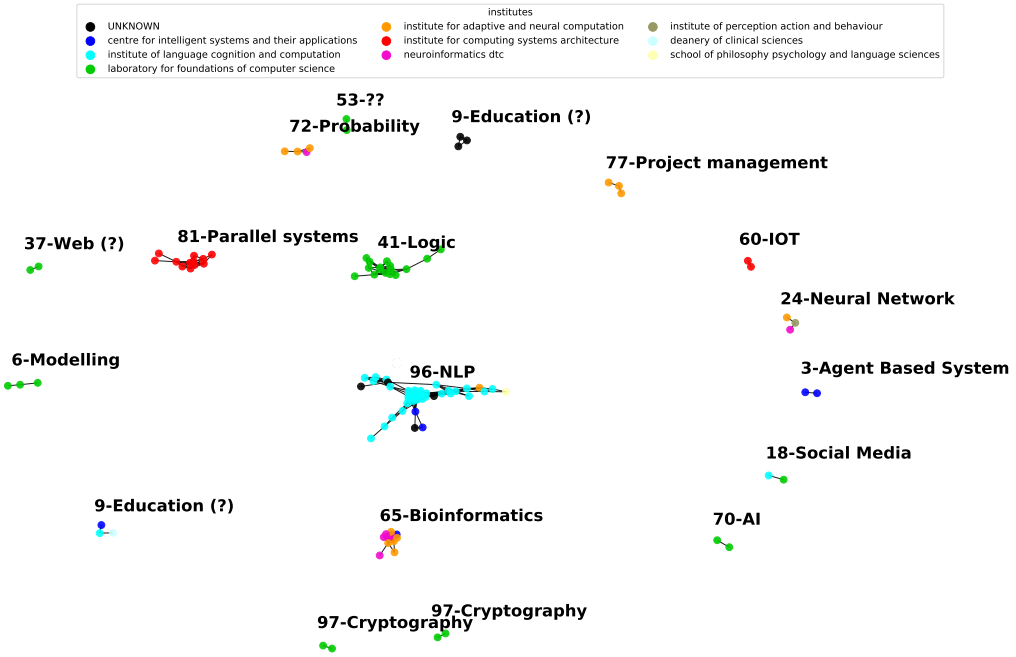


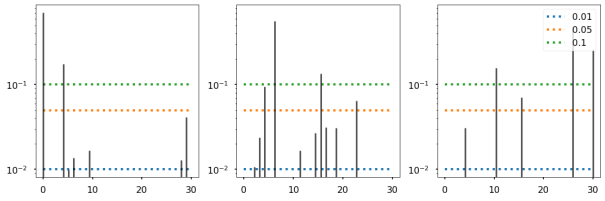
Figure 5.2: We created `topicnetref-6yr`, using collection of publication's metadata from dblp from 1997 to 2017. The resulting network can be seen as placing the publications from the school in context in the larger computer science community. We see multiple topic-based communities similar to those in `topicnet-6yr`(Figure 5.1), such as Bioinformatics and a more general cluster on Natural Language Processing (NLP) that would encompass Computational Linguistics and Speech Synthesis.

by the Law of Large Number, bias in distribution of papers from each topic would not affect the topic model derived. However, this intuition should be used with caution, as Computer Science is a fast moving field and some topics may be a field of study since its inception, while others may be a more recent field. On hindsight, we should have used a collection from 2012 to 2017 instead of the entire collection from 1997-2017, in tandem with the collaboration network used. Overall, using a reference corpus provided an alternate view of the topics in network, and is useful to discover topic-based clusters especially when the publications used is limited or have a skewed distribution.

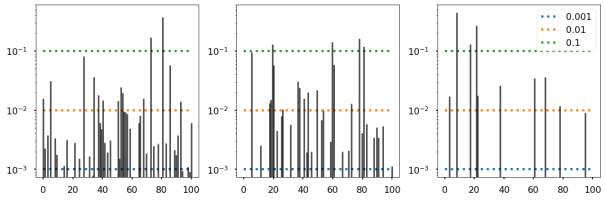
5.3 Discussion

5.3.1 Subtle differences in inputs

If we consider the collection of publication that went into the each model - collaboration network, topic modelling, and deriving topic-similarity network- we will realise a subtle difference between the inputs. Using the same set of publication, topic modelling sees all the documents; collaboration network only takes into account documents with two or more collaborators from Informatics; to infer topic distribution for each



(a) Minimum probability used for topic-similarity networks derived with `tm-20yr` or `tm-6yr` is 0.01, which is sufficient for salient topics to be captured.



(b) We use minimum probability of 0.001 topic-similarity networks derived with `tm-dblp`, since there is a total of 100 hidden topics to be discovered.

Figure 5.3: We randomly sampled three different researchers to generate underlying distribution of hidden topics. Different values of minimum probabilities can affect the topic proportions used for calculating the similarity between researchers. This is a risky move as a topic that is close to the minimum probability will be masked as 0, which have downstream effect when calculating threshold over the similarity matrix.

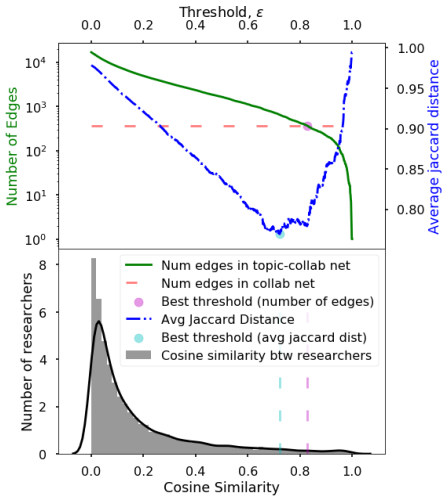


Figure 5.4: We search across possible values of ε as a cut-off for the distribution of cosine similarity between researchers (below). Two approach is proposed, one using global descriptive statistics of the collaboration network (`infnet-6yr`) - the number of edges, while the other uses average jaccard distance between researchers in the topic network and the ground-truth collaboration network. We see that thresholds for each criterion is different, suggestive that the networks have different properties. This will be evaluated in the next Chapter Communities and social influence.

researcher in informatics, the topic model takes in all documents that the researcher participated in. This means that the the topic similarity network takes into account more documents than is being used to generate the collaboration network. The difference is extremely prominent when a researcher's set of publications contains work that are participated outside of the School.[1] We think that this difference would not lead to major differences because an author's field of research may switch between subfields, but the underlying topic distribution would not see much shift, since our topic model have seen the entire collection of publications. We rest our case on the assumed robustness of topic model.

5.3.2 Noisy and irrelevant topics

From Chapter 4, we saw that some topics does not have any meaning (such as topics describing the published `Springer`), or are incoherent (low C_V score). Intuitively, we should also ignore the probabilities derived from these topics, so that our similarity measure between researchers account take these noise into account. Our approach to

[1] As noted in Chapter 2, this could also refer to individuals who were once in the School and have had left.

these topics is to set the minimum probability on the values returned. Figure 5.3 illustrate the effect of different minimum values, where the probability for topics that are lower than the minimum are reduced to 0. This means that when calculating similarity f, we only take into account the interests that are prominent in topics that are highly probable.

Our method so far is simplistic, slightly more sophisticated method exists. One could use a *weighted similarity measure* for derived topic probability, where weights for topics that are irrelevant/incoherent are set to 0. Another possible approach is to use the coherence score derived as the weights, so that incoherent topics are unrepresented. We did not implement this in this project, and leave this as future work.

Chapter 6

Communities and social influence

The final stroke of our exploratory work involves comparing networks we have created so far. The difference between each network is the set of linkages between researchers. We saw previously that topic-similarity networks connect researchers based on how similar their topic distributions are; collaboration network connects researchers based on presence of work done together; weighted versions of each network characterise these connections by providing a numerical quantity to strength of ties. Specifically, we are interested in how these networks differ from each other using principle of homophily. Using a node attribute, such as institute membership, do members of an institute collaborate with each other more than with others outside the institute?

In addition, we explore communities underlying each networks using two algorithms: 1) Girvan-Newman algorithm, a divisive algorithm that iteratively remove edges that are high in edge betweenness; 2) Modularity maximisation, an agglomerative algorithm that iteratively connect edges between nodes if the additional linkages increases the modularity. In terms of collaboration network, the first algorithm remove edges where most number of shortest path goes through, which usually corresponds to connections between individuals from two different social group. On the other hand, the second algorithm builds a network based on the idea of modularity - a measure of quality of communities compared to a network that is connected to each other at random. Networks with high modularity hence have communities whose connections are unexpected of a random network. We are interested in community detection in two ways: 1) Communities discovered may provide new insight into relationships between researchers; 2) A topic-similarity network is an instance of topical communities - which means we should compare it against another network where communities are obvious for fairness.

This chapter hence investigates communities in network and homophily. First, in Section 6.1 we elaborate our testing methods, highlighting parameters and their meaning. Then, we outline the communities detected in our network in Section 6.2 and results from homophily test in Section 6.3. Finally, in Section 6.4 we set our work in context of existing literature and provide exploratory work in future.

6.1 Methodology

We compare the communities detected from the network against some ground-truth labels as well as test for homophily behaviour (homophily test).

Communities We can compare communities in Informatics Network as out dataset contains institute membership information of researchers .[1] This would correspond to our postulation that network structure reflects communities; communities reflects working group; working groups are the institutes that individuals that are part of.

To that end, we use two different community detection algorithms:

1. Girvan-Newman edge betweenness algorithm (GN) - a divisive hierarchical algorithm that iteratively remove edges where most number of shortest path passes through - the "arbitrator" of information (Girvan and Newman, 2002). In collaboration networks, edge with high betweenness may represent bottlenecks or bridges of information where pairs of researchers each from different fields collaborated on an interdisciplinary paper.

2. Modularity maximisation with Louvain heuristics (Mod) aims to maximise modularity Q. While directly maximising modularity is computationally hard, approximation algorithms exists.[2] We use an implementation that starts with as many communities as nodes and iteratively move a node to its neighbourś community if the gain in modularity is positive (Blondel et al., 2008).

These algorithms are implemented in NetworkX and Community packages respectively (Hagberg et al., 2008; Aynaud, 2018). For GN, we created our own function to choose the best partition that maximises the modularity. Concretely, we calculate and store the modularity of each partition at every level of the divisive algorithm and pick the best partition that gives the highest modularity.

Homophily test Suppose an undirected network with individuals from two institutes a and b, and probability of observing each institute is p and q respectively, equivalent to the fraction of individuals in the network. Then we would observe cross-institutes edges if the first end of edge is a and the other end is b, or vice versa. This means that the probability of observing cross-institute edges is $2pq$ if edges were formed randomly independent of institutes. Hence, homophily behaviour exists if fraction of cross-institute edges is significantly less than $2pq$ (Easley and Kleinberg, 2010, Chp 4.).

Considering seven distinct types for an attribute, say institute, we compare homophily in two ways:

[1] We illustrated in Chapter 2 that institute membership information is incomplete where only 94.3% of researchers who have a publication, being listed in one of the seven schools in Informatics.

[2] Modularity $Q \in [0,1]$ where 1 indicates strong community structure, where strong means linkages within a community is higher than across the network where edges are connected randomly. A Q value of 0 indicates number of intra-community edges is no better than random. Typically, Q falls between 0.3 and 0.7 (Newman and Girvan, 2003).

Networks	Hypothesis
`infnet-6yr` *vs* `topicnet-6yr`	Both networks should exhibit similar structure since edges in topic-similarity network connotates the effect of collaboration.
`infnet-6yr` *vs* `infnet-6yr(w)`	Weighted edges symbolise collaborative relationship between pair of nodes, and `infnet-6yr(w)` would be a more accurate model than a unweighted network.
`topicnet-6yr` *vs* `topicnetref-6yr`	Using dblp (reference corpus) to derive a similarity network,the topic similarity should be generalised. Does `topicnetref-6yr` provides a more accurate model of collaborative relationship?

Table 6.1: We focus on collaboration networks derived from collections from 2012-2017 because it is more complete compared to the collection from 1997-2017. Each hypothesis highlight the intended effect we wish to observe when comparing two networks.

1. Pairwise comparison between each attribute type; total of $\binom{7}{2} = 21$ comparisons. Homophily exists if two institutes collaborate more within each institute than between institutes by only accounting for edges involving nodes and edges with either of these attributes.

2. Comparison of each type against the rest of the network; total of seven comparison. This considers if a particular institute collaborates more outside of the institute.

Homophily attributes On top of common (professional) interests, such as *institute membership*, we also test if individuals who are popular are more likely to work with one another using *node degrees* to connotate "popularity" in a network.[3]

Networks Across the previous two chapters, we have developed multiple versions of `infnet` and topic-similarity network (Table 5.1). In the interest of space, we limit our discussion to networks derived using publications from 2012-2017 as the quality of data is objectively better. Table 6.1 provides the networks we are pitting against each other and the hypotheses for comparing them.

6.2 Communities in Informatics

This section investigates communities detected in collaboration network - `infnet-6yr` and `infnet-6yr(w)` where difference between the two networks is the presence of weighted edges capture additional qualitative relationships that simple networks are unable to. Our community detection algorithms GN and Mod derived different sets of communities (Figure 6.1). By holding the position of nodes constant across the

[3] Another thought is to use clustering coefficient (Watts and Strogatz, 1998) - the fraction of possible triangles for each node. However, it is undefined if the node is isolated or a leaf; effectively measuring only nodes with degrees ≥ 2. This limits the ability to capture homophily across the network.

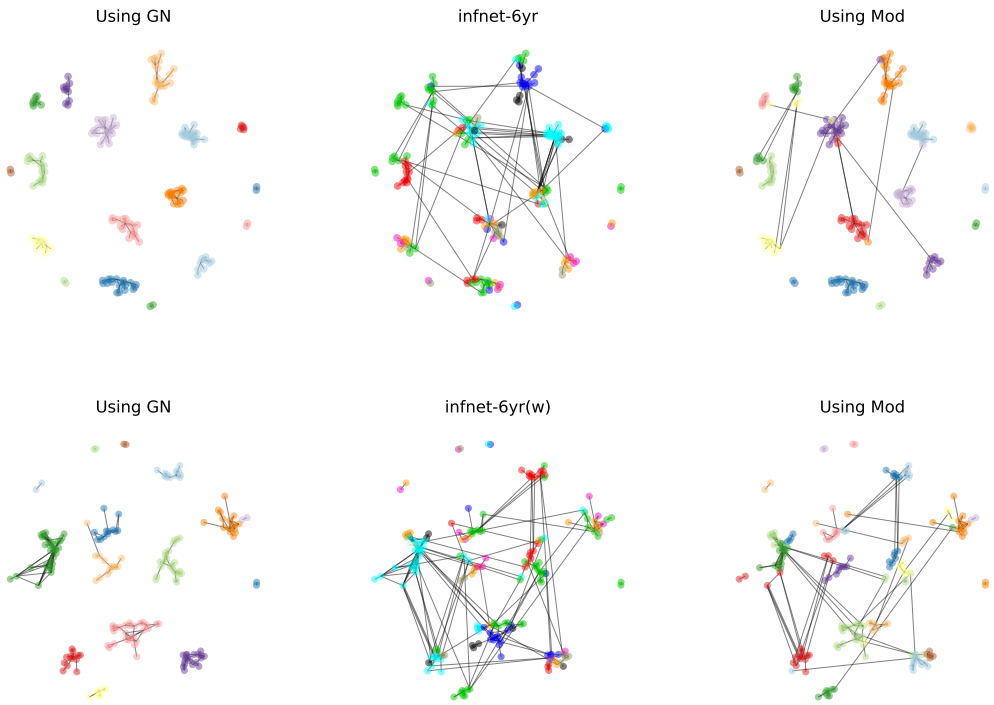


Figure 6.1: We compare the communities derived from Girvan-Newman algorithm (left column) and modularity maximisation (right column), for `infnet-6yr` and its weighted version - `infnet-6yr(w)`. We fix the layout to the communities detected in `infnet-6yr` using Girvan-Newman algorithm (top left). This allow us to visually see the edges that are removed from the original network located in the centre, and contrast both clustering algorithms. Apart from the middle column where nodes are coloured according to the institute they represent, the nodes are coloured according to the community they fall in.

layouts, we are able to visualise the edges that were being removed in order to derive each community.

Our collaboration network `infnet-6yr` have six connected components; the giant connected component (GCC) accounts for 94.6% of all the nodes. Both clustering algorithm, further sub-divided the GCC into 10 components. Interestingly, applying Mod on `infnet-6yr(w)` resulted in the GCC being divided into 16 components instead. In general, both clustering algorithm perform differently, where Mod chose to reserve some edges that GN decide to remove. If we compare the best modularity score Q, both clustering algorithm perform the same for `infnet-6yr`(Mod: 0.76, GN: 0.75), but with weighted edges, Mod have a higher modularity score than GN (0.82 vs 0.74). This could suggest that the clustering found by Mod is 'better'.

Topic-based communities found in topic-similarity network We attempted to use GN and Mod on topic-similarity networks, and found that the original network itself have high modularity, and both algorithms did not further sub-divide the networks. We postulate that this is because by the process of creating topic-similarity networks, topic-based communities were already discovered. Our topic-similarity networks, unlike collaboration networks, do not have a GCC despite having roughly the same number of edges. Hence, clustering on topic-similarity network is futile.

		Difference between fraction of cross-institute edges and expected probability $2pq$			
Institute	**Institute size**	`infnet-6yr`	w/ Mod	`topicnet-6yr`	dblp corpus
CISA	18	**-0.110**	**-0.129**	**-0.101**	-0.051
IANC	20	**-0.283**	**-0.292**	**-0.305**	**-0.374**
ICSA	21	-0.064	-0.054	-0.036	**-0.117**
ILCC	42	**-0.155**	**-0.164**	**-0.179**	**-0.208**
IPAB	12	-0.067	-0.058	***0.024***	-0.044
LFCS	49	-0.064	-0.071	-0.061	-0.011
Neuro	12	-0.008	-0.008	-0.005	-0.014

Table 6.2: The more negative the difference is from the expected probability $2pq$, the more significant the institute exhibit homophily. We found that IANC exhibits homophily by institute the most, while *Neuroinformatics DTC* (Neuro) the least. Using dblp corpus for topic-similarity network exaggerates some homophily behaviour of some institute (e.g. ILCC and IANC) but the reverse is observed for LFCS. This may be due to the high similarity score between publications by members of ILCC and IANC, resulting in more intra-institute edges being formed. As observed in Figure 5.1, members from LCFS have publications from different domains such as Logic and Cryptography and is separated into different communities. Although these topic-based communities are formed largely by individuals from the same institute, the larger and more tightly connected community ILCC results in higher homophily test score.

6.3 Homophily - Does similarity beget authorship?

Now, we investigate if researchers collaborate with each other based on institute membership and popularity (node degree). We apply homophily test on four models: 1) `infnet-6yr`, 2) `infnet-6yr(w)` and its variant with clusters found using Mod; 3) `topicnet-6yr`, and 4) `topicnetref-6yr`.

6.3.1 Institute membership

In the first case where we compare each institute in turn against the rest of the institutes, we found evidence that suggests that nodes have preference of collaborating within institute compared to across institute (Figure 6.2). This behaviour is more apparent for some institutes: Institute of Adaptive and Neural Computation (IANC) and Institute of Language, Cognition and Computation (ILCC) which exhibited strong evidence of homophily by institute across all models.

Comparing different models, we saw some interesting observations:

1. Conducting homophily test with the network split into communities discovered provides stronger evidence that institutes have homophily behaviour, as fraction of cross-institute edges decreases. This is true for most institutes, except *Neuroinformatics DTC*. On closer inspection, we observe that there are no collaboration work between members of the institute.

2. Testing homophily on topic-similarity network with respect to institute shows

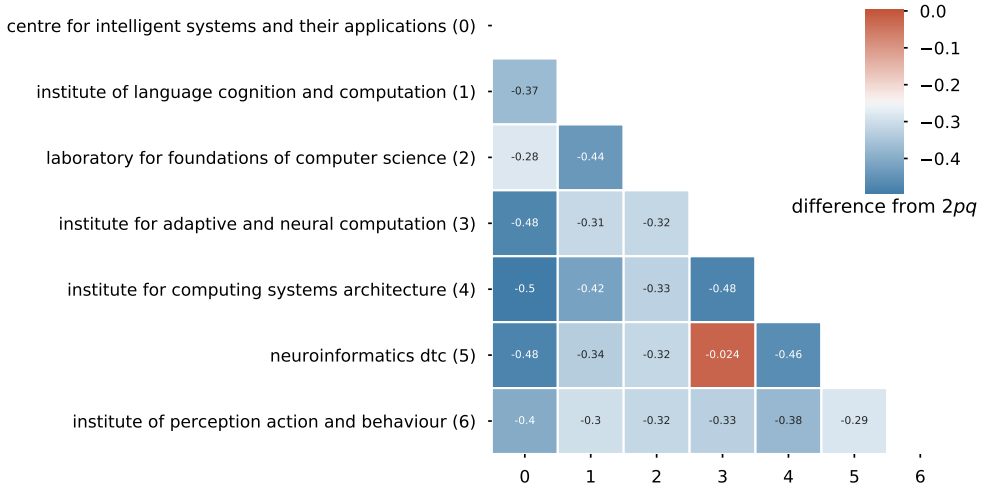
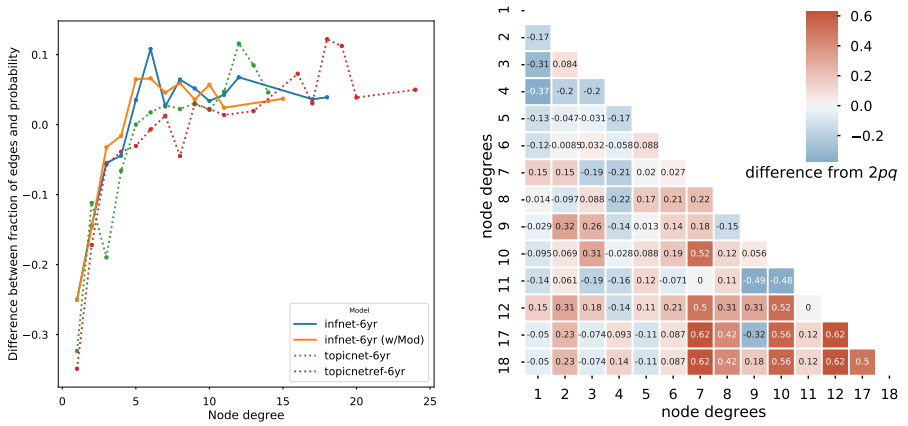


Figure 6.2: We compare homophily test between institutes in Informatics. Once again, a positive score suggests inverse homophily while a negative score from the expected number of cross-institute edges suggests homophily by institute. We observe that between *neuroinformatics DTC* and *Institute of Adaptive and Neural Computation*, cross-collaboration is more prominent than other institute pairs (red square). A darker blue square is suggestive of a homophily by institute where members prefer to collaborate more than random within institute.

that topic-based communities decreases the number of edges between institutes for most institutes when compared with `infnet-6yr` (w/Mod). This provide some evidence that topic-based communities are prototype of communities. However, we would be cautious to make any strong claim, as looking at the institute size, we would expect that larger institutes would have more publications. Consequently, when our trained topic model infer topic probabilities, papers that are more well-represented in the corpus would have less ambiguity, leading to strong topic similarity score. These are the edges that will usually be preserved and hence remains in the network.

3. If we use a larger corpus with the aim to prevent these skewed distribution for highlighting the differences, our homophily test based on institutes for `topicnetref-6yr` provides stronger evidence of homophily for most institutes. Interestingly. without the use of the dblp corpus, IPAB have a positive value, suggesting that it might exhibit some inverse homophily - that members of the institute prefer to collaborate outside of the institute. However, when used with a dblp corpus, the difference deceased by 3 times. We think that the publications by members of IPAB may be strongly under represented in the topic model trained on publications from *Edinburgh Research Explorer*.

We also conducted a homophily test between pair of institutes by calculating the fraction of edges going between the two institutes compared to the total number of edges between and within each of the two (Figure 6.2). We observe that between *neuroinformatics DTC* and *Institute of Adaptive and Neural Computation*, compared to other institutes it collaborates with, provides evidence that cross-collaboration is more prominent between this two institutes. This observation is seen in other networks derived



(a) Case 1: Comparing each node degree against other degree

(b) Case 2: Comparing each pair of node degrees

Figure 6.3: Our analysis on the homophily by networks for cases shows that nodes with low degree exhibit homophily behaviour; while node with high degree do not but prefer to collaborate with others. Our analysis, however is limited for large node degree since these corresponds to a small number of individuals.

too. Apart from the above observation, we think that it is limited to have any other claims about relationship between institutes, because in the actual network, there are other institutes that are present and members of each institute is not limited to the other institute to collaborate with.

6.3.2 Node degree

We investigate if nodes with high degree prefer to associate with others with high degree too. In social networks, we can imagine two types of people - extroverts who are very outgoing and make friends with many others , or introverts who prefer to keep to themselves. Sociological literature highlight that extroverts have a tendency to make friends with other extroverts, and likewise for introverts. Such "preferential attachment" based on one's character led to high degree nodes (extroverts) making connections with nodes that are of high degree too, resulting in degree correlation in the network (Newman, 2003).

In the previous section, we saw that collaboration within institutes are more likely - some more than others. This shows that our network is more of an affiliation network - where members are part of a group and hence would most likely work with others within the group - than a social network. Following this line of reasoning, if the group is large, then one would form multiple connection within the group leading to high node degree. Conversely, members belonging to smaller groups would have lower node degree.

For case one, comparing difference between fraction of edges going between nodes of different degree and the expected probability, we found evidence that low degree

nodes exhibit homophily, while inverse homophily is present for high node degree (Figure 6.3a). This is not surprising since nodes with low node degree are by definition connected to lesser nodes; coupled with the presence of leafs (nodes with degree of 1) in the network, nodes with low node degree would be attached more to nodes with other lower degree nodes. For nodes with high degree, we observe inverse homophily instead, which means that there are less edges going between high degree nodes than we should expect. We interpret this result in twofold:

1. The number of high degree node (more than 10) is small between 2 to 3 each. Hence, probability $2pq \rightarrow 2p$ and is a smaller number. This means that our homophily test would result in inverse homophily as long as edges there exists edges between nodes of other degree. This account for the high positive difference we see;

2. From the collaboration network point of view, high degree nodes are usually individuals who have been in the School for a substantially long period. Indeed, we observe that 'reader', 'personal chair,' 'chair,' 'research fellow', and 'lecturer' are commonplace. This means that they these individual would usually be a mentor to project group lead of a research group in the institute where group members are usually PhD students. This means that we would see connections between high degree nodes, lower degree nodes, explain our observation of inverse homophily for high degree nodes.

For case two we conduct pairwise homophily test between two different node degrees. and it appears clear that nodes with lower node degree, some level of homophily is observed, while between nodes with high degree, inverse homophily is observed (Figure 6.3b). From both cases, we observe that homophily between nodes with low degree, and the inverse for node with high degree. Lastly, comparing the models, not much difference can be observed between them.

6.4 Concluding remarks

This project is an exploratory work on collaboration network and similarity networks based on topic discovered using topic modelling - topic-similarity network. We chose to explore bibliometric materials from *Edinburgh Research Explorer*, and show that we can accurately create collaboration network describing relationship between researchers in School of Informatics. Making use of the dataset, we created topic models using LDA, and use it to discover topic similarity between individuals. To create a topic-similarity network, we use collaboration network to guide the derivation of topic-similarity network. We further explored communities in both networks with the aim of investigate relationships between researchers in the School.

Using topic-similarity networks, we discovered topic-based communities in Informatics which are labelled communities using common topics each researcher in the community is interested in. We postulate that this is a case of clustering, and think that

it could be used for context dependent soft-clustering too. These topic-similarity networks also have the potential to be used for exploring like-minded individuals in Informatics - spurring collaboration in the School.

We use the networks created to investigate the the underlying hypothesis of our project - that both collaboration network and topic-similarity network should exhibit *similar* structure and cohesion for individuals in the network. We defined similarity using the principle of homophily - that individuals of the same characteristics tends to form ties. We show that this is the case for the institute structure of the School where individuals are allocated into of the seven institute. Also, contrary to usual social networks where nodes with low degree tends to form ties together and likewise for high degree nodes, we observed inverse homophily for the case of high degree nodes. We think that this is because a high degree node is a more senior researcher in the School or a project lead. This means that they would participate nurturing duties by assisting researchers who have just joined the school.

Our work so far is largely in tandem with publications on scientific collaboration network, and have only explored a miniute space of this subject. Further work on this project includes:

- Exploring collaboration across the University, instead of a School
- Providing methods qualify nodes that are connected to members of the institute through an external member
- Exploring the impact of information flow within a network - similar to diffusion of ideas or innovation in Social Network

Bibliography

Aletras, N. and Stevenson, M. (2013). Evaluating Topic Coherence Using Distributional Semantics. *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*, pages 13–22.

Amaral, L. a. N., Scala, A., Barthlmy, M., and Stanley, H. E. (2000). Classes of small-world networks. *Proceedings of the National Academy of Sciences*, 97(21):11149–11152.

Arun, R., Suresh, V., Madhavan, C. E. V., and Murthy, M. N. N. (2010). On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. In *Advances in Knowledge Discovery and Data Mining*, Lecture Notes in Computer Science, pages 391–402. Springer, Berlin, Heidelberg.

Aynaud, T. (2018). python-louvain: Louvain Community Detection. original-date: 2016-08-16T14:18:17Z.

Barabsi, A.-L. and Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512.

Barabsi, A. L., Jeong, H., Nda, Z., Ravasz, E., Schubert, A., and Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3):590–614.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77.

Blei, D. M. and Lafferty, J. D. (2007). A correlated topic model of Science. *The Annals of Applied Statistics*, 1(1):17–35.

Blei, D. M. and Lafferty, J. D. (2009). Topic models. *Text mining: classification, clustering, and applications*, 10(71):34.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Brinkmeier, M. and Schank, T. (2005). Network Statistics. In *Network Analysis*, Lecture Notes in Computer Science, pages 293–317. Springer, Berlin, Heidelberg.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., and Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296.

Chuang, J., Manning, C. D., and Heer, J. (2012). Termite: visualization techniques for assessing textual topic models. In *In Proceedings of the International Working Conference on Advanced Visual Interfaces*, page 74. ACM Press.

Easley, D. and Kleinberg, J. (2010). *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press. Google-Books-ID: atfCl2agdi8C.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3-5):75–174. arXiv: 0906.0612.

Franceschet, M. (2010). Collaboration in computer science: a network science approach. Part I. *arXiv:1010.4747 [physics]*. arXiv: 1010.4747.

Franceschet, M. (2011). Collaboration in computer science: a network science approach. Part II. *arXiv:1104.4296 [physics]*. arXiv: 1104.4296.

Freeman, L. C., Roeder, D., and Mulholland, R. R. (1979). Centrality in social networks: ii. experimental results. *Social Networks*, 2(2):119–141.

Girvan, M. and Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826.

Goffman, C. (1969). And What Is Your Erdos Number? *The American Mathematical Monthly*, 76(7):791.

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

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235.

Grossman, J. W. (2002). The Evolution of the Mathematical Research Collaboration Graph. page 12.

Hagberg, A., Swart, P., and S Chult, D. (2008). Exploring network structure, dynamics, and function using NetworkX. Technical Report, Los Alamos National Lab (LANL), Los Alamos, NM (United States).

Hoffman, M., Bach, F. R., and Blei, D. M. (2010). Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pages 856–864.

Jeong, H., Neda, Z., and Barabasi, A.-L. (2003). Measuring preferential attachment for evolving networks. *Europhysics Letters (EPL)*, 61(4):567–572. arXiv: cond-mat/0104131.

Kaiser, M. (2008). Mean clustering coefficients: the role of isolated nodes and leafs on clustering measures for small-world networks. *New Journal of Physics*, 10(8):083042. arXiv: 0802.2512.

Kleinberg, J. (2000). The Small-world Phenomenon: An Algorithmic Perspective. In *Proceedings of the Thirty-second Annual ACM Symposium on Theory of Computing*, STOC '00, pages 163–170, New York, NY, USA. ACM.

Li, W. and McCallum, A. (2008). Pachinko allocation: Scalable mixture models of topic correlations. *J. of Machine Learning Research. Submitted*.

Medina, A., Matta, I., and Byers, J. (2000). On the origin of power laws in Internet topologies. *ACM SIGCOMM Computer Communication Review*, 30(2):18.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., and McCallum, A. (2011). Optimizing Semantic Coherence in Topic Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Minka, T. P. and Lafferty, J. (2012). Expectation-Propogation for the Generative Aspect Model. *arXiv:1301.0588 [cs, stat]*. arXiv: 1301.0588.

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108. Association for Computational Linguistics.

Newman, M. E. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the national academy of sciences*, 101(suppl 1):5200–5205.

Newman, M. E. J. (2001a). Clustering and preferential attachment in growing networks. *Physical Review E*, 64(2).

Newman, M. E. J. (2001b). Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E*, 64(1).

Newman, M. E. J. (2001c). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E*, 64(1).

Newman, M. E. J. (2003). Mixing patterns in networks. *Physical Review E*, 67(2). arXiv: cond-mat/0209450.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582. arXiv: physics/0602124.

Newman, M. E. J. (2008). Mathematics of Networks. In *The New Palgrave Dictionary of Economics*, pages 4059–4064. Palgrave Macmillan, London.

Newman, M. E. J. (2016). Community detection in networks: Modularity optimization and maximum likelihood are equivalent. *arXiv:1606.02319 [cs.SI]*.

Newman, M. E. J. and Girvan, M. (2003). Finding and evaluating community structure in networks. *Physical review E*, 74(3):036104.

Newman, M. E. J., Watts, D. J., and Strogatz, S. H. (2002). Random graph models of social networks. *Proceedings of the National Academy of Sciences*, 99(suppl 1):2566–2572.

Ng, A. Y., Jordan, M. I., and Weiss, Y. (2001). On Spectral Clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, pages 849–856. MIT Press.

Rehurek, R. and Sojka, P. (2010). Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer.

Roder, M., Both, A., and Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408. ACM Press.

scrapy, d. (2018). scrapy: Scrapy, a fast high-level web crawling & scraping framework for Python. original-date: 2010-02-22T02:01:14Z.

Sievert, C. and Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pages 63–70. Association for Computational Linguistics.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., and Su, Z. (2008). Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 990–998. ACM.

Teh, Y. W., Newman, D., and Welling, M. (2007). A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation:. Technical report, Defense Technical Information Center, Fort Belvoir, VA.

Tomassini, M. and Luthi, L. (2007). Empirical analysis of the evolution of a scientific collaboration network. *Physica A: Statistical Mechanics and its Applications*, 385(2):750–764.

Travers, J. and Milgram, S. (1969). An Experimental Study of the Small World Problem. *Sociometry*, 32(4):425–443.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of small-world networks. *Nature*, 393(6684):440.

Yusuke, S. (2018). pdfminer.six: Python PDF Parser. original-date: 2014-08-29T14:04:53Z.