Dynamic Analysis of Social Networks

Freddy Bruce

Christ Church

University of Oxford

Supervisor:

Prof. Stephen Roberts

Trinity 2017

FINAL HONOUR SCHOOL OF
ENG / EEM (delete as appropriate)

DECLARATION OF AUTHORSHIP

You should complete this certificate. It should be bound into your fourth year project report, immediately after your title page. Three copies of the report should be submitted to the Chairman of examiners for your Honour School, c/o Clerk of the Schools, examination Schools, High Street, Oxford.

Name (in capitals): ……………………………………………………………………………………………

College (in capitals): …………………………………… **Supervisor:** …………………………………………

Title of project (in capitals): ………………………………..…………………………………………………

Page count (excluding risk and COSHH assessments): ………………………………………………………

Please tick to confirm the following:

I have read and understood the University’s disciplinary regulations concerning conduct in examinations and, in particular, the regulations on plagiarism (*The University Student Handbook. The Proctors’ and Assessors’ Memorandum, Section 8.8*; available at https://www.ox.ac.uk/students/academic/student-handbook) □

I have read and understood the Education Committee’s information and guidance on academic good practice and plagiarism at https://www.ox.ac.uk/students/academic/guidance/skills. □

The project report I am submitting is entirely my own work except where otherwise indicated. □

It has not been submitted, either partially or in full, for another Honour School or qualification of this University (except where the Special Regulations for the subject permit this), or for a qualification at any other institution. □

I have clearly indicated the presence of all material I have quoted from other sources, including any diagrams, charts, tables or graphs. □

I have clearly indicated the presence of all paraphrased material with appropriate references. □

I have acknowledged appropriately any assistance I have received in addition to that provided by my supervisor. □

I have not copied from the work of any other candidate. □

I have not used the services of any agency providing specimen, model or ghostwritten work in the preparation of this project report. (See also section 2.4 of Statute XI on University Discipline under which members of the University are prohibited from providing material of this nature for candidates in examinations at this University or elsewhere: http://www.admin.ox.ac.uk/statutes/352-051a.shtml.) □

The project report does not exceed 50 pages (including all diagrams, photographs, references and appendices). □

I agree to retain an electronic copy of this work until the publication of my final examination result, except where submission in hand-written format is permitted. □

I agree to make any such electronic copy available to the examiners should it be necessary to confirm my word count or to check for plagiarism. □

Candidate’s signature: …………………………………………….. **Date:** ………………………..

Acknowledgements

This project would not have been possible without the excellent help and advice of Prof. Stephen Roberts, which has been invaluable.

Abstract

Traditionally, research on network theory has focused on static data, yet most real networks are naturally dynamic. Social network analysis is becoming increasingly significant in fields including sociology and recommender systems. The present work is motivated by the Enron Corpus a large scale real email dataset, relating to the bankruptcy of Enron in 2001. Based on this dataset, we provide a collection of statistical and computational methods for modelling and exploring the social structure of directed communication networks.

This report proposes a technique for mapping the temporal data to a series of directed social networks using modularity. Given the extracted graphs, a method for obtaining the community structure from directed graphs is proposed — therefore detecting information flow. Finally, these methods are applied to the Enron Corpus to show how these techniques can be used to reveal the most influential nodes and to track communities through time.

Contents

List of Figures

Chapter 1

Introduction

1.0.1 Motivation

Traditionally, research on network theory has focused on static data, yet most real networks are naturally dynamic. Social network analysis is becoming increasingly significant in fields including sociology and recommender systems. The present work is motivated by the Enron Corpus a large scale real email dataset, relating to the bankruptcy of Enron in 2001. Based on this dataset, we provide a collection of statistical and computational methods for modelling and exploring the social structure of directed communication networks.

In the next section, we present an overview of the current report structure and provide a description of each chapter.

1.0.2 Overview

In **Chapter 2**, we present a review of key elements of the social network paradigm and introduce important terminology that is used throughout the report. We focus on real world networks' topological properties; small world effect, heavy tailed distributions and community structure.

In **Chapter 3**, we introduce the Enron Corpus and explore the timeline of events that led to its bankruptcy in 2001. Following this, we present a short review of existing research on the dataset and provide an initial analysis.

Following our discussion on network analysis and the Enron Corpus, we proceed in **Chapter 4** to discuss the task of inferring graph structure from temporal data streams.

We discuss how modularity can be used to determine time window size for extracting the maximally informative graph.

In **Chapter 5**, we focus on detecting communities in a directed network. We present a review of non-negative matrix factorisation for community detection and explain how this can be extended to directed networks. The method also addresses issues of overlapping communities and defining node membership scores.

In **Chapter 6**, we apply our network analysis techniques to the Enron email dataset and explore the social network across a two year span. We show how various graph measures relate to the state of the company and investigate their dynamics over the entire timespan.

Finally, we conclude in **Chapter 7** providing an overview of our findings, both from a methodological and social perspective.

Chapter 2

Network Analysis

The study of networks has been an important field of research since Euler and the solution of the Königsberg bridge problem (1735).

During the 20th century, the network paradigm was used to model human interactions with Milgrams Small-world experiment [Milgram, 1967] and Granovetters Strength of weak ties [Granovetter, 1977] being the most famous examples. Given its generality, network analysis has undergone surging growth in recent years [Buchanan and Caldarelli, 2010]-with a broad range of researchers exploring methods for defining network topologies and underlying properties such as community structure, vulnerability and diffusion potential.

Computational advances have also allowed researchers to study large-scale systems such as animal societies, social media and the World Wide Web [Newman, 2015]. In addition, Open-source network analysis toolboxes like *NetworkX* [Hagberg et al., 2008] and *Gephi* [Bastian et al., 2009] provide standard algorithms and visualisation packages supporting additional investigation. The theoretical study of networks has introduced new methods for representing time-evolving phenomena [Porter and Gleeson, 2016], and furthermore how networks may be manipulated to inform recommendations or predictions. This latter element of exploration has been of importance to machine learning researchers, as the occurrence of network structures in real-world (particularly social) settings has given rise to new models for community detection and edge prediction within incomplete or noisy datasets.

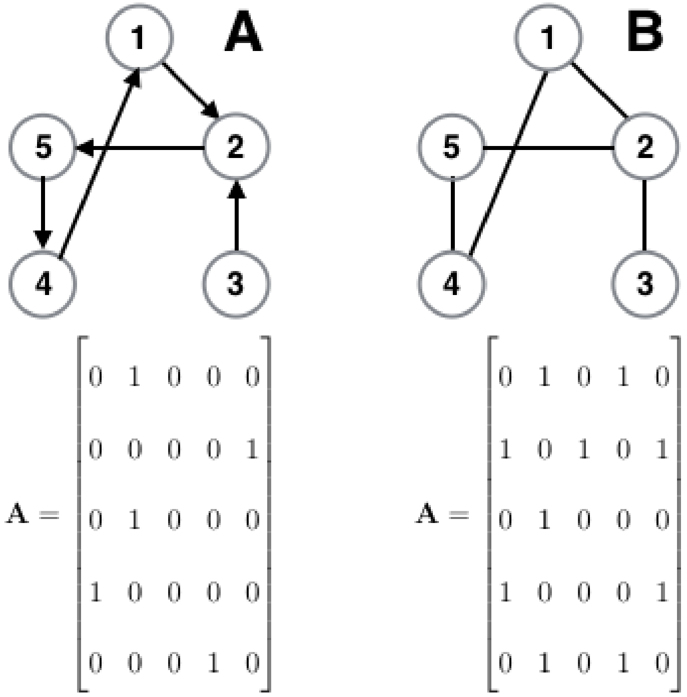
In this section, we present the fundamentals of network analysis, following which we describe the properties of real world networks, including heavy-tailed distributions and the small-world effect and, finally we investigate community structure.

2.1 Graphs and Networks

In mathematics, a network is called a graph and therefore these terms may be used interchangeably. A graph $\mathcal{G}$ consists of a set of nodes $\mathcal{V} = \{1, , N\}$ and a set of edges $\mathcal{E} = \{(i, j), i, j \in \mathcal{V}\}$, i.e. $\mathcal{E}$ consists of pairs of vertices in $\mathcal{V}$, with the understanding that $(i, j) \in \mathcal{E}$ means that there is an edge from node i to node j. In an undirected graph, if (i, j) is an edge, so is (j, i) whereas for a directed graph it may be the case that there is an edge from i to j but no edge j to i. The adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ can be used to describe the connectivity pattern, so that if $a_{ij} \neq 0$ then nodes i and j are linked together; a_{ij} can take a Boolean value (unweighted edge) or a real value (weighted edge). For the purpose of this study, we look at the graph across the time axis and assume a series of discrete timepoints $t_1, ..., t_n$, therefore at timepoint t_i we observe the graph instance $\mathcal{G}(\mathcal{V}_i, \mathcal{E}_i)$.

The term network is used to describe the interactions in a system, where nodes are individual entities and edges represent some form of association. Networks make it possible to characterise the complex systems of our world, in the same way that a map describes the surrounding landscape [Rosvall, 2006].

The exponentially increasing popularity of network analysis in scientific literature [Buchanan and Caldarelli, 2010] has been triggered by two reasons. Firstly, the computational advances in recent decades in data gathering, storage and processing [Rosvall, 2006]; but also, the recognition that real-world systems are comprised of many entities interacting such that their collective behaviour is not a simple combination of their individual behaviours. The traditional reductionist viewpoint is to break a system down into its individual parts and analyse each part individually [Rosvall, 2006]. The network paradigm excludes the specific characteristics of each node and instead describes that data as a series of relationships. This makes it an appropriate tool for describing systems on a macroscopic level, as nodes and edges can describe any associations between entities. In



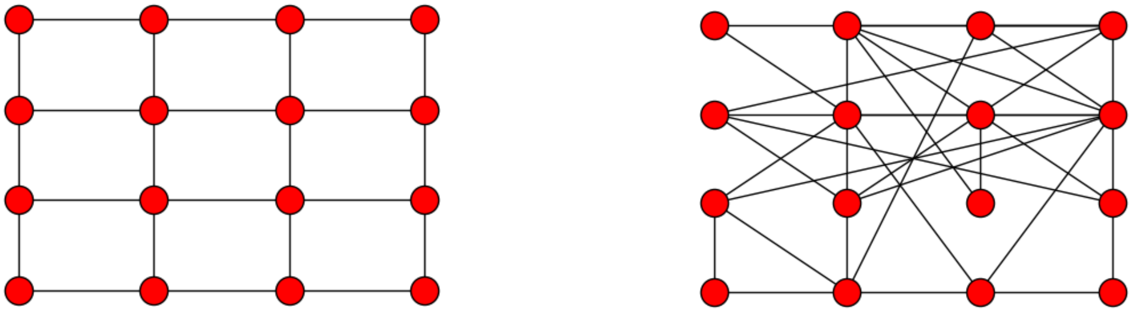
$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \qquad \mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \end{bmatrix}$$

Figure 2.1: Two simple graphs and their adjacency matrices. Graph A is directed and graph B is undirected.

conclusion, the fundamental idea of the paradigm is that the connectivity pattern has a significant effect on the behaviour of the overall system [Newman, 2015].

2.2 Real World Networks

Networks allow us to model real complex systems, for example, economic networks where companies are the nodes, and the links symbolise the purchases and sales or financial loaning, and the weight of the links represent the value of these transactions. Intuitively, it is easy to conclude that not all nodes are connected and that connections do not appear at random. The connectivity pattern of a real-world network is not an ordered lattices, nor completely random [Watts and Strogatz, 1998], but a structure that emerges from the self-organisation mechanisms of the individual nodes. Properties such as heavy-tailed distribution, small-world effect and community structure [Newman, 2003] can explain the formation of the network. We will, for this reason discuss heavy-tail distribution and small-world effect in Subsections 2.2.1 and 2.2.2. Due to its importance to the report, we describe community structure in greater detail in Section 2.3.



(a) Ordered lattice.

(b) Random Graph.

Figure 2.2: Subfigure 2.2a shows a regular grid of 25 nodes. Subfigure 2.2b shows an Erdős-Rényi (ER) random graph. These graphs were generated using the *Networkx* package [Hagberg et al., 2008] for Python.

2.2.1 Heavy-Tailed Distribution

The degree of a node is the number of nodes connected to it. For a given adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$, we define the degree distribution as $d_i = \sum_{j=1}^{N} a_{ij}$. The degree distribution is the fraction of nodes with degree d. [Albert et al., 1999] showed that for real world networks the degree distribution deviates from the Poisson form of Erdős-Rényi random graphs [Newman et al., 2001]. Instead, they exhibit a heavy tail. The authors [Albert et al., 1999, Newman, 2015] showed that the degree distribution follows a power law.

$$p(d) \propto d^{-\gamma}, \text{for } d > d_{min}, \tag{2.1}$$

where γ is a heuristic dependent constant that is generally $2 < \gamma < 3$.

Heavy tailed distributions have reported in many complex systems. The most famous being the measuring of the diameter of the World Wide Web [Albert et al., 1999].

2.2.2 Small-World Effect

Stanley Milgram's "Small World Experiment" was a historical study in the field of network analysis, in which letters were passed between individuals in order to reach a desired target. The experiment showed that it only required 6 steps [Milgram, 1967].

In the context of graphs, we define the geodesic distance g_{ij} as the number of edges between node i and node j. A network exhibits small world properties if the geodesic distance scales logarithmically of slower with network size for fixed mean degree [Newman, 2003].

Real world networks exhibit transistivity, or large clustering coefficient, therefore, if A and B are connected to C then it is likely that A is connected to B. This triangle formation increases the different ways to reach a node and results in a low geodesic distance. This means real world networks can exhibit "Ultrasmall Worldness" [Cohen and Havlin, 2003].

2.3 Community Structure

Identifying and evaluating community structure is an active area of research. Community structure is the way in which a given network is clustered into several latent (and possibly overlapping) classes of nodes - creating "hot-spots" of increased connectivity [Psorakis, 2013]. Early community detection methods were based on graph partitioning methods but were limited by the requirement that the number of communities must be specified. Agglomerative methods were used, but these tended to neglect periphery nodes [Newman and Girvan, 2004]. The seminal work of Mark Newman and Michelle Girvan [Newman and Girvan, 2004] proposed a class of divisive methods which removed edges based on a range edge related metrics. This work created an explosion of new algorithms with the focus being on latent feature models.

There are several ways in which a community can be defined. It can be described by structure, for example, "cliques" or by similarity of nodes, where members who have similar semantics are assumed to be similar. Further, we can label communities as either crisp (a node belongs to exactly one community at a time) as in [Aggarwal and Yu, 2005] or soft (a node belongs to a community with some probability or possibility) as in [Sarkar and Moore, 2005]. The underlying model of computation usually constrains the definition: the communities are crisp, if stream clustering is used; the communities are fuzzy, if fuzzy clustering is used; if dynamic probabilistic modelling is used then the latent variables are communities and these latent variables that describe each data instance with a specific likelihood, thus allowing an instance to be a member of many communities [Spiliopoulou, 2011].

Chapter 3

Enron Corpus

The Enron corpus is a large email dataset which was released for research purposes following the Federal Energy Regulatory Commission's investigation into the Enron. The original dataset was made available to the public and searchable via the web; however, due to the volume of emails over 160GB, this made it impractical to use. Subsequently, copies of the collected emails were made available on hard drive media. William Cohen from CMU also released a copy of the dataset online for researchers (http://www-2.cs.cmu.edu/enron).

The email corpus is appealing to researchers as it is a large-scale email collection of a real organisation over 3.5 years. The dataset is of particular interest for social network analysis because it enables the long-term examination of interactions among the entities of an organisation.

This section provides a synopsis of the Enron case, following which we describe the refined database. Finally, we provide initial analysis results and as a result develop of our research questions.

3.1 Enron Timeline

Enron formed in 1985 through the merger of Houston National Gas (a utility company) and Internorth of Omaha (a pipeline company) under the direction of Kenneth Lay. It grew to become the 7th biggest company in revenue for 15 years through their activity of buying electricity from generators and selling to consumers. The deregulation of the energy markets allowed Enron to proclaim themselves as an energy broker. They identified areas where capacity was less than the demand, built power plants and then sold them before they lost value [Diesner et al., 2005]. As Enron's energy business flourished, they applied their brokering skills to markets such as bandwidth and television advertising time. By 2002, they were employing 21,000 people in 40 countries [BBC, 2002]. Enron hired Arthur Andersen LLP as their auditors.

In 1999, Enron began separating their losses from equity and derivative trades into special purpose entities (SPE) and these partnerships were excluded from the company's net income. An example of an SPE was Raptor, where a group of Enron Executives used loaned stock money from Enron and bought equity shares in two companies, Avici and New Power Co. Enron then profited from the increase in the value of the SPE. Raptor then took the losses and thus removing them from their financial reports. This resulted in an off-balance sheet financing system [Diesner et al., 2005].

In December 2000, President and Chief Operating Officer (COO), Jeffrey Skilling took over from Kenneth Lay as CEO. Lay remained as Chairman, and the company's stock hit a high of $84.87. However, by August 2001, Skilling surprisingly resigned, and Lay was named as CEO again [Diesner et al., 2005]. In the same month, Sherron Watkins Enrons VP of Corporate Development, became a whistleblower, writing an anonymous letter to Lay accusing Enron of inappropriate use of SPEs and fraud [Diesner et al., 2005].

In October 2001, the losses transferred to SPEs totalled $618 million and were publically reported in the third quarter [Diesner et al., 2005]. Shortly after this announcement the SEC commenced an enquiry into Enron and revealed that the total losses for the last five years totalled $568 million. In November 2001, Lay resigned, and shares valued at less than $1 [Diesner et al., 2005].

Andersen had information regarding the financial situation at Enron, before the formal insolvency but did not make it public [Diesner et al., 2005] and jointly Andersen and Enron deliberately classified hundreds of millions of dollars of shareholders' equity as an increase rather than a decrease. In 2000, Andersen's internal senior management had also rated Enron lower than they evaluated the client to the public. In October 2002, Enron was requested by Andersen to destroy any documentation relating them to Enron.

CEO Kenneth Lay, CFO Andrew Fastow and his lieutenant Michael Kopper appeared before Congress in February 2002, and all of them pleaded the Fifth Amendment [BBC, 2002]. Former Enron Treasurer, Ben Glisan Jr, was the first employee to be imprisoned after pleading guilty to conspiracy in September 2003 [Diesner et al., 2005]. In January 2004, Fastow pleaded guilty, and the FBI charged his wife, Lea and seven Executives. Finally in February, Skilling was accused of fraud, conspiracy, filing false statements to auditors and insider trading.

3.2 Enron Data

The Federal Energy Regulatory Commission (FERC) first posted the corpus in 2002 to provide an explanation as to why the FERC was investigating Enron. The corpus contained 619,449 emails from 158 employees. Each email provides information on the sender, receiver, sent timestamp, subject, body and text. However, this dataset had integrity problems. Leslie Kaelbing from MIT had purchased the dataset which a group at SRI International collected and prepared the data for the CALO (Cognitive Assistant that Learns and Organizes) project [Calo, 2002]. The team corrected the original integrity problems and made the dataset further available.

In March 2004, William Cohen from CMU also put the data online. This version of the dataset contains 517,431 emails from 151 users in 3500 folders. Each message contains both the sender and receiver's email address, date and time, subject, body, text and some other email specific technical details [Diesner et al., 2005]. However, it also contains all kind of emails personal and official. Some of the emails have, however, been deleted as part of the redaction effort due to requests from affected employees.

The version of the dataset used in this report is provided by Jitesh Shetty and Jafar Adibi [Shetty and Adibi, 2004]. This dataset was primarily cleansed by removing blank, duplicated and junk emails, and as a result the dataset totals 252,759 emails from 151 employees. Shetty and Adibi migrated the dataset into a MySQL database which contained tables for the employees, messages, recipients and reference information. We are using this version of the dataset as it has a very well documented cleaning process, and the structure of the MySQL database make it advantageous.

3.2.1 Initial Analysis

Figure 3.1 shows the total number of emails sent each month by individuals through the corpus. There are some peaks in the communications shown, which relate to events in the organisation. The highest peaks are October and November 2001 when the crisis broke out, and the investigation began. The low points are January and February, and August and September which are vacation periods.

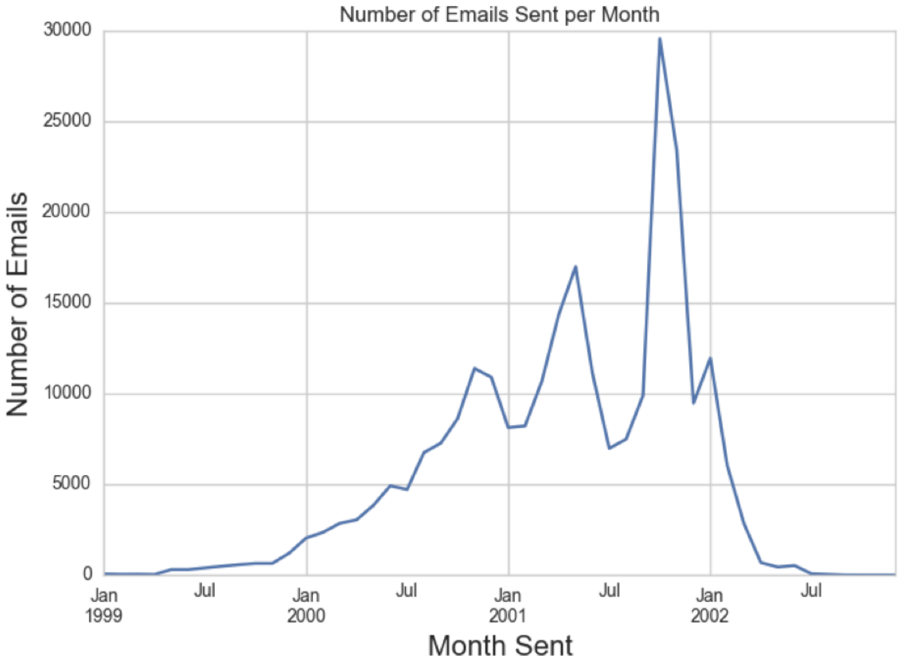


Figure 3.1: Number of emails sent by the 151 employees.

Figure 3.2 shows the number of emails sent by individuals. The figure suggests that a minority sent the majority of emails.

To come to an understanding of the relationship between positions during different situations at the company, we compiled Table 3.1. This table shows the emails exchanged for each position for the months of October 2000 and October 2001. It indicates that generally, people receive more emails than they send which agrees with the heavy-tailed

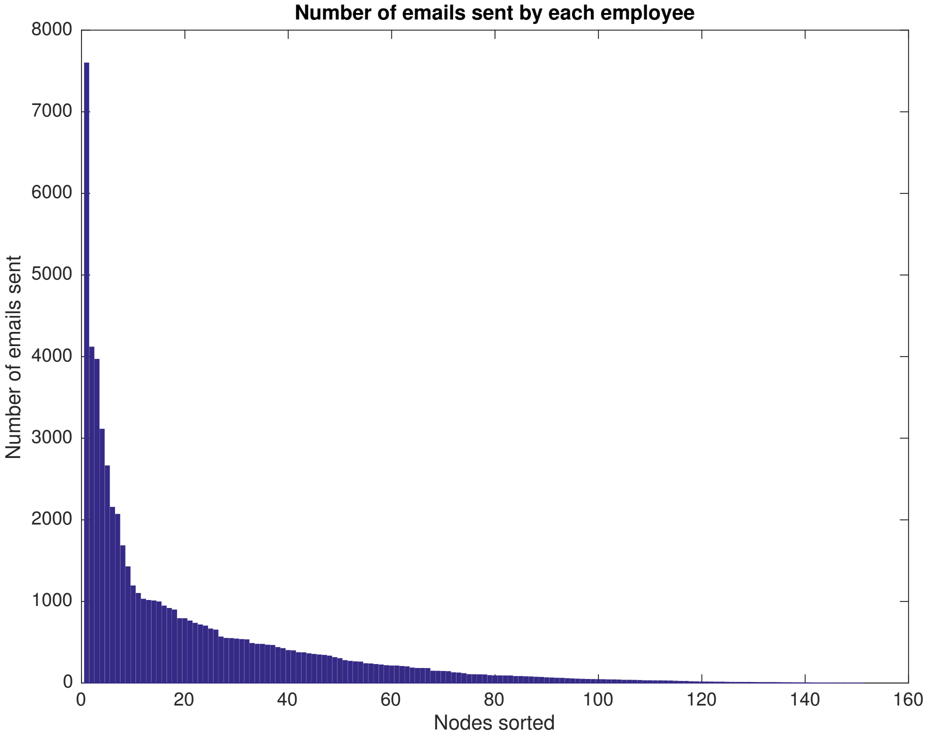


Figure 3.2: Number of emails sent by each employee across the entire dataset. It suggests that a minority sent the majority of emails.

distribution indicated in Figure 3.2. However, it also shows that in 2001 higher ranks were consuming more (receiving more emails than sending) than the year 2000.

[Diesner et al., 2005] ran graph level centrality measures to represent the heterogeneity or dispersion of agents in the Enron network. They found that the network is less centralised that other networks, suggesting a highly segmented workforce with little cross communication. However, after the crisis broke the network became more centralised—suggesting the inequality of importance, cross communication and group cohesion increased.

3.3 Discussion

In this chapter, we have discussed the Enron Corpus and the timeline of events that led to the downfall of Enron; in addition to an initial analysis of the dataset. The majority of research conducted on the Enron Corpus focused on Natural Language Processing (NLP) [Klimt and Yang, 2004b, Klimt and Yang, 2004a]. There has been some

Table 3.1: Emails exchanged in the months of October for the years 2000 and 2001.

Position	Sent 2000	Received 2000	Sent 2001	Received 2001
CEO	24.7%	75.3%	1.0%	99.0%
President	57.8%	42.2%	23.3%	76.7%
Vice President	21.3%	78.7%	15.4%	84.6%
Man. Dir.	28.4%	71.6%	2.9%	97.1%
Director	8.0%	92.0%	10.9%	89.1%
Manager	36.3%	63.7%	19.1%	80.9%
Trader	48.0%	52.0%	15.4%	84.6%

research from a social network perspective, with [Shetty and Adibi, 2004] generating a social network for the 151 employees and providing information on quantitative features. [Diesner et al., 2005] used social network analytic techniques to the exploration of structural and behavioural features of the organisation. Finally, there has been work exploring the community structure of the corpus [Qian et al., 2006]. On the basis of this and the preliminary analysis, we seek to further our analysis, from a group of descriptive statistics to a *Social Network Analysis* perspective—the advantages of which are discussed in Section 2.1. The preliminary exploration also highlighted that the dataset is inherently time-related and directed in nature, features previously not explored. As a result, our dataset poses a series of motivating questions:

- How can we infer a social network from a relational database containing transactional data, at the same time accounting for the temporal nature of the data?
- As groups and cliques are common in a social context and workplace context, how can we discover communities within a given directed communication social network?
- How can we track individuals within the *Social Network Analysis* perspective?

By addressing the above questions, we seek to reveal how the social network changes across two years and how these properties and entities relate to various phases of a company's life cycle. The insights gained by the analysis we perform and propose are of potential further benefit for modelling the development of crisis scenarios in organisations and the investigation of indicators of failure.

Chapter 4

Inferring Graph Structure from Temporal Data

The objective of a stream mining algorithm is to maintain an up-to-date model of the data [Spiliopoulou, 2011]. Therefore, the learned model must adapt to the visible records currently. In some cases, this is all the records seen, however, for many types of analysis, such as identification of influential nodes or tracking the growth and shrinkage of communities, it is reasonable to discard some graph elements [Spiliopoulou, 2011]. In stream mining, this is implemented with a sliding window in which only data within the window are considered by the model. Furthermore, the choice of the length of the window has consequences on the characteristics of the resulting networks [Holme, 2003]. For example, short windows it is seen that different behavioural patterns play a role during weekends and weekdays, and for longer windows it is seen that networks aggregated during holiday periods are significantly different.

In this chapter, we provide a description of our data format, we present a common approach of discovering network structure from temporal data, which is based on a discretisation of the observation stream given an appropriate resolution parameter.To conclude, we look at how to calculate this resolution parameter.

4.1 Description of Data

As mentioned in Section 3, the Enron dataset is a large-scale email collection of a real organisation consisting of approximately 252,000 emails over 3.5 years. The data generated

Table 4.1: Sample format of data.

Sender ID	Recipient ID	timestamp
50	6	2000-04-04 10:35:00
50	147	2000-04-04 10:35:00
76	50	2000-04-04 10:40:00
23	69	2000-04-05 02:20:00
48	147	2000-04-05 02:27:00

is essentially a stream of timestamped transactions, a sample can be seen in Table 4.1.

Let our temporal data stream $\mathcal{D}$ be represented in the form:

$$\mathcal{D} = \{\mathrm{ID}_{sender}, \mathrm{ID}_{receiver}, t_z\}_{z=1}^{Z}$$

where Z is the number of records in our database. Taking a single data point can be read as: Observation $\#$ z node: ID_{sender} sent an email to node: $\mathrm{ID}_{receiver}$ at time t_z. Note that t_z represents an event time, so $t_z - t_{z+1}$ is not constant. The timestamps have also been converted to seconds for convenient manipulation. The data can be visualised like Figure 4.1, where the horizontal axis is time and each data record is represented by a stem. This visualisation will be used frequently in this section to explain further concepts.

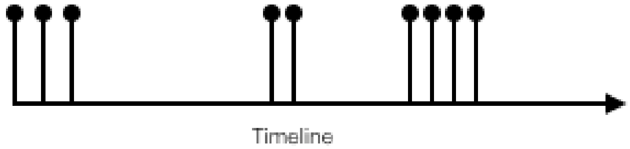


Figure 4.1: Example data stream. The horizontal axis is time and each data record is represented by a stem.

Finally, the data can be considered to just be a transactions table in a relational database. Therefore, restricting our analysis to metrics such as total number of emails sent per person or the distribution of emails through time. The goal of this section is, therefore, to find an appropriate mapping of the data stream $\mathcal{D}$ to adjacency matrices $\mathbf{A} \in \mathbb{R}^{N \times N}$, where a_{ij} represents the extent of social affiliation of person i to person j.

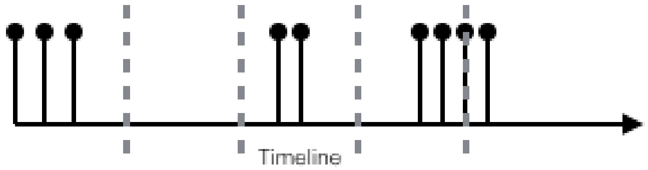


Figure 4.2: An example data stream discretised into intervals of fixed length Δt.

4.2 Fixed Time Window

Within the Enron data set, our core assumption is that communication between two individuals suggests a social interaction while the more communication between two individuals the stronger the social tie. Based on this, a fixed time window approach is employed which involves placing a link between two individuals i, j, if they are observed in the data stream within a fixed temporal distance Δt [Krings et al., 2012]. See Figure 4.2, that shows a snapshot of the data $\mathcal{D}$, that has been discretised into a series of intervals of fixed length Δt. Within each bin we place a link between nodes and so output at directed weighted adjacency matrix **A**, where a_{ij} is the number of times node i has communicated with node j.

If the data $\mathcal{D}$ collection period is of length T, the total number of intervals is $n = \frac{T}{\Delta t}$. The algorithm performs a linear search at each interval, $\mathcal{O}(n)$. Then within in each interval the algorithm performs a linear search to construct the adjacency matrix. Within each time window, the number of unique individuals is N and the number of occurrences of each individual can vary from 0 to Z. Therefore, the computational cost of the process is $\mathcal{O}(ZNn)$, where Z and n are competitive terms. If n becomes large, the time window becomes small and so typically the number of observation Z decreases. When n becomes small, the time window increases and so the number of observations also increases. The size of the fixed time Δt has a significant effect on the topology of the network, which will be tackled in Section 4.4. Furthermore, the position of the interval boundaries can also affect the network structure. For example, two communications may appear in close temporal proximity but have an interval boundary between them (see the last two intervals of Figure 4.2). This sort of structure may lead us to miss important connections in the

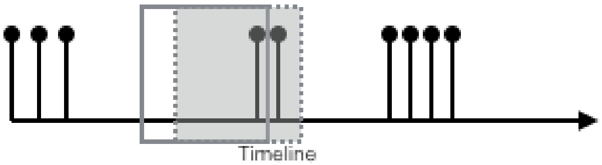


Figure 4.3: An example data stream showing a rolling fixed time window.

data stream. In the next section, we introduce an extension to the fixed time window to overcome this issue.

4.3 Sliding Window

The position of the interval boundaries can also affect the network structure. For example, two communications may appear in close temporal proximity but have an interval boundary between them (see Figure 4.2). This sort of structure may lead us to miss important connections in the data stream, and in order to overcome this, we can employ a sliding window, where we increment the window forward a small amount. This means the community detection model which will be discussed in Section 5 will be adapting from records only within the sliding window. This will allow us to track the most influential nodes and understand how communities are changing. For example, new nodes may join the network or a node may exhibit changes in their properties or preferences. There are several elaborate extensions to this method that assign weights to the records inside the time window [Nasraoui et al., 2003] and filling the time window with the expected most representative records [Bifet and Gavald, 2007].

4.4 Time Window Size

In the time window method, the selection of the parameter of Δt is crucial to the extracted network topology, as the interactions a_{ij} are a function of this scale parameter, and therefore, too small a time window, may lead to missing important co-occurrences; whilst too large a time will lead to an overestimation of the connectivity.

In situations where we have no prior knowledge of the scale parameter Δt, we have to examine multiple time windows against a performance metric. [Krings et al. 2012] have experimented with network metrics, such as clustering, network density and modularity — the latter we use as a coefficient.

4.4.1 Modularity

Suppose that we have a network in which the nodes are classified by some characteristic that has a finite set of possible values, for example, the nodes could be people, and the characteristics could be gender, race or nationality. We can measure the assortativity of the network, to find the fraction of edges that run between vertices of the same type (community), and then subtract the fraction of such edges that we would expect if the network were random. For the trivial case, when all the nodes are of the same type (or community), all the vertices will be connected. The randomised expected network will be the same as there is nowhere else the edges can fall. The difference of the two numbers is then zero, telling us that there is no non-trivial assortativity in this case [Newman, 2015]. Alternatively, the network has no community structure. There is community structure when there are more edges between nodes of the same type (or community) that we would expect by chance.

In mathematical terms, let us denote the community or type of vertex i by c_i, where $i \in 1, ..., n$, n being the total number of communities. The total number of edges that run between nodes in the same community is:

$$\sum_{\text{edges } ij} \delta(c_i, c_j) = \frac{1}{2} \sum_{ij} A_{ij} \delta(c_i, c_j), \tag{4.1}$$

where $\delta(c_i, c_j)$ is the Kronecker delta.

To calculate the randomised network, consider an edge attached to vertex i, which has degree k_i. Each edge has two ends in the network, therefore there are $2m$ ends of edges, where m is the total number of edges. If connections are made purely at random, then the probability that the other end of an edge is one of the k_j ends attached to vertex j is thus $\frac{k_j}{2m}$ [Newman, 2015]. Counting all k_i edges attached to i, the total expected number of edges between vertices i and j is then $\frac{k_i k_j}{2m}$, and the expected number of edges between all pairs of vertices of the same type is:

$$\frac{1}{2}\sum_{ij}\frac{k_i k_j}{2m}\delta(c_i, c_j) \tag{4.2}$$

Taking the difference of Equation 4.1 and Equation 4.2 gives the difference between the actual and expected number of edges:

$$\frac{1}{2}\sum_{ij} A_{ij}\delta(c_i, c_j) - \frac{1}{2}\sum_{ij}\frac{k_i k_j}{2m}\delta(c_i, c_j) = \frac{1}{2}\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(c_i, c_j) \tag{4.3}$$

Conventionally, one calculates the fraction of such edges, which is given by Equation 4.3 divided by the total number of edges m:

$$Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(c_i, c_j) \tag{4.4}$$

This quantity is called *Modularity*. We can see from this definition that when the fraction of edges in a community is the same as the fraction of edges connecting to a node within that community, the modularity Q is 0 and, therefore, there is no community structure. Q is driven to 1 when there is strong community structure thus we see a significant fraction of edges within communities rather than between communities. Most real-world networks have modularity values ranging from 0.3 to 0.7 [Newman and Girvan, 2004]. Weighted networks can use modularity [Newman, 2004], where the null model, rather than being defined by degree sequence, is defined by strength sequence.

4.4.2 Calculating the Enron Corpus Window

The time window size has been calculated using modularity as a performance metric. The maximum modularity is calculated for time window sizes from 1 day to 2 years with 1 day increments. The results are shown in Figure 4.4. On average there are 84 emails sent per day and so in that initial period we have very sparse adjacency matrices and as a result can get very high modularity values. The peak at 6 days is most likely due to the small number of data points and so is ignored. The maximum value of Q is 0.8034 and is for a time window of size 103 days. This window size will be used as the size of the rolling window.

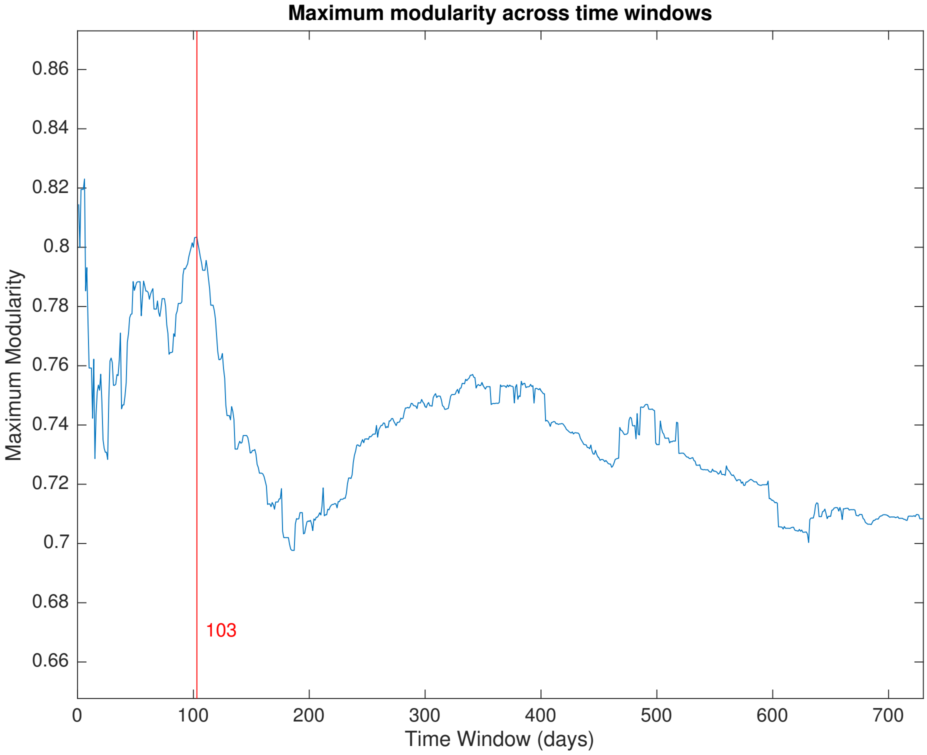


Figure 4.4: We plot the modularity metric Q for 730 time window sizes starting at 1 day and finishing at 730 days (2 years). The peak at 6 days we regard due to the small number of data point. The maximum value of Q is 0.8034 and is for a time window of size 103 days.

4.5 Discussion

In this section, we focussed on the problem of constructing a social network from a temporal data stream. Real data streams have no ground truth graphs, we are aiming to discover the maximally informative network structure from which the observed data is to rise. By discretising the stream based on some fixed time window Δt and drawing links between them. We used modularity to calculate Δt and calculated a window size of 103 days.

Chapter 5

Overlapping Community Detection via Non-Negative Matrix Factorisation

5.1 Introduction

In this chapter, we described a novel approach by [Psorakis et al., 2011] to community detection that adopts a Bayesian non-negative matrix factorisation model to achieve soft partitioning of a network. We explain how community detection can be viewed as a generative model in a probabilistic framework with priors existing over the model parameters. To conclude, we outline how this approach can be input with asymmetric adjacency matrices (directed graphs), and therefore, communities are no longer comprised of mutually connected nodes but correspond to groups where members point to a common target.

5.2 Model Formulation

5.2.1 Background

Community structure can be viewed as an underlying mechanism that drives social tie formation. For example, the Enron Corporation is a large institution which consists of many departments. The employees can participate across many departments, and this participation has a strong effect on their social circle. As a result, the key modelling assumption is that co-participation of two individuals across a given range of communities increases the likelihood that those individuals are interacting [Psorakis, 2013].

To encode these ideas in a mathematical framework, we consider an example network of N web pages connected via hyperlinks. We assume that this network of websites can be organised into C thematic groups that correspond to overlapping, topic-focused user communities, e.g. sport, politics, etc. Given a topic c, a certain page i can either serve c with intensity h_{ci}, or seek c with intensity w_{ic} [Psorakis, 2013]. Therefore the association strength $\hat{a}_{ij}$ is based on their thematic compatibility which is:

$$\hat{a}_{ij} = \varphi(\mathbf{w}_i, \mathbf{h}_j), \tag{5.1}$$

where φ is a given compatibility function. $\mathbf{w}_i, \mathbf{h}_j$ are C-dimensional vectors which encode the "seek/serve" community membership powers. We consider φ to be a linear function:

$$\hat{a}_{ij} = \mathbf{w}_i^T \mathbf{h}_j, \tag{5.2}$$

so we can express the network adjacency matrix via the following factorisation:

$$\hat{\mathbf{A}} = \mathbf{W}\mathbf{H}, \tag{5.3}$$

where **A** $\in \mathbb{R}^{N \times N}$, **W** $\in \mathbb{R}^{N \times C}$ and **H** $\in \mathbb{R}^{C \times N}$.

Based on the above, we pose the community detection problem in the following manner — "if connection weight among individuals in a network results from the degree of homophily in their community membership profiles, which membership structure is the most plausible for explaining the observed connections?" [Psorakis, 2013] Alternatively, given an adjacency matrix $\mathbf{A}$, which factorisation $\hat{\mathbf{A}} = \mathbf{W}\mathbf{H}$ approximates A as well as possible given a particular distance metric?

5.2.2 Generative Model

Figure 5.1 shows the graphical model. As explained in subsection 5.2.1, a_{ij} represents the count of interactions between individuals i and j in the directed weighted network **A** $\in \mathbb{R}^{N \times N}$. We assume that there are C 'hidden' classes of nodes that affect a_{ij}. We can, therefore, classify groups of nodes to communities as latent variables that allow us to explain the increased interaction density in specific regions of the network [Psorakis et al., 2011].

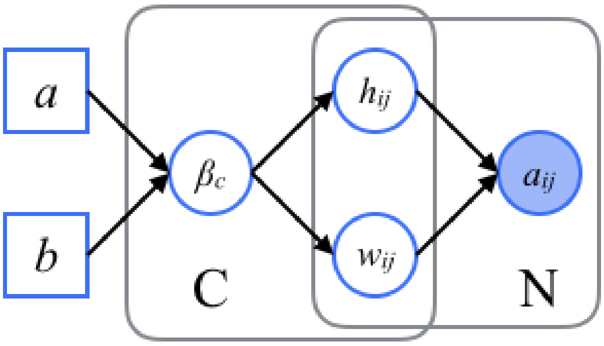


Figure 5.1: Graphical model showing the generation of count processes **A** from the latent structure **W** and **H**.

We can assume that an unobserved expectation network $\hat{\mathbf{A}}$ influence the interactions in the network **A**. The elements of $\hat{\mathbf{A}}$ are the expected number of interaction between individuals i and j. The expectation network can, therefore, be written as $\hat{\mathbf{A}} = \hat{\mathbf{W}}\hat{\mathbf{H}}$, where $\hat{\mathbf{W}} \in \mathbb{R}^{N \times C}$ and $\hat{\mathbf{H}} \in \mathbb{R}^{C \times N}$. So, we model each interaction a_{ij} as being drawn from a Poisson distribution with rate $\hat{a}_{ij} = \sum_{c=1}^{C} w_{ic} h_{cj}$. C represents the number of unknown communities and each element $c \in 1, ..., C$ in the row i of **W** and column j accounts for the contribution of a single latent community.

The latent variables w_{ic}, h_{cj} have shrinkage or automatic relevance determination priors [Mackay, 1995] with scale hyperparameters $\boldsymbol{\beta} = \{\beta_c\}$ [Tan and Fevotte, 2013]. This allows the appropriate model order to arise from a single run. We start with a large C (the maximum number being the number of nodes), the priors moderate complexity by 'shrinking' close to zero irrelevant columns of **W** and rows of **H** that are not explaining **A**. This is performed by placing distributions that expectations approach zero unless non-zero data is input. The distribution of β_c is parameterised by a and b.

Considering the graphical model in Figure 5.1, we can write the joint distribution as:

$$p(\mathbf{A},\mathbf{W},\mathbf{H},\boldsymbol{\beta}) = p(\mathbf{A}|\ \mathbf{W},\mathbf{H}) p\mathbf{W}|\boldsymbol{\beta}) p(\mathbf{H}|\boldsymbol{\beta}) p(\boldsymbol{\beta}) \tag{5.4}$$

Therefore the posterior over model parameters is:

$$p(\mathbf{W},\mathbf{H},\boldsymbol{\beta}|\mathbf{A}) = \frac{p(\mathbf{A}|\ \mathbf{W},\mathbf{H})p(\mathbf{W}|\boldsymbol{\beta})p(\mathbf{H}|\boldsymbol{\beta})p(\boldsymbol{\beta})}{p(\mathbf{A})} \tag{5.5}$$

5.2.3 Posterior-based cost function

The aim is to maximise the model posterior given the observation, or, alternatively, to minimise the negative log posterior which is an energy (or error) function $\mathcal{U}$. Noting that $p(\mathbf{A})$ is constant on inference over the model's free parameters, we can define:

$$\mathcal{U} = -\log p(\mathbf{A}|\ \mathbf{W},\mathbf{H}) - \log p(\mathbf{W}|\boldsymbol{\beta}) - \log p(\mathbf{H}|\boldsymbol{\beta}) - \log p(\boldsymbol{\beta}) \tag{5.6}$$

The first term is the log-likelihood of our data; we can derive this from the probability $p(\mathbf{A}|\ \mathbf{W},\mathbf{H}) = p(\mathbf{A}|\hat{\mathbf{A}})$ of observing each interaction a_{ij} given a Poisson rate $\hat{a}_{ij}$. So, stating the negative log-likelihood of a single observation as:

$$-\log p(a|\hat{a}) = -a\log\hat{a} + \hat{a} + \log a! \tag{5.7}$$

using the Stirling approximation to second order:

$$\log a! \approx a\log a - a + \frac{1}{2}\log 2\pi a \tag{5.8}$$

Equation 5.7 can be written as:

$$-\log p(a/\hat{a}) \approx a\log\big(\frac{a}{\hat{a}}\big) + \hat{a} - a + \frac{1}{2}\log 2\pi a \tag{5.9}$$

Finally, for all observed data the negative log-likelihood is:

$$-\log p(\mathbf{A}|\hat{\mathbf{A}}) = -\sum_{i=1}^{N}\sum_{j=1}^{N}\log p(a_{ij}|\hat{a}_{ij}) \simeq \sum_{i=1}^{N}\sum_{j=1}^{N}\left(a_{ij}\log\frac{v_{ij}}{\hat{v}_{ij}} + \hat{v_{ij}} + \frac{1}{2}\log(2\pi v_{ij})\right) + \kappa \tag{5.10}$$

where κ is a constant.

Over the columns of **W** and rows of **H**, we place independent half normal priors with precision parameters $\boldsymbol{\beta} \in \mathbb{R}^C = [\beta_1, ..., \beta_C]$. The negative log-likelihood of the priors over **W** and **H** are defined as:

$$-\log p(\mathbf{W}|\boldsymbol{\beta}) = -\sum_{i=1}^{N}\sum_{c=1}^{C}\log \mathcal{HN}(0,\beta_c^{-1}) = -\sum_{i=1}^{N}\sum_{c=1}^{C}\left(\frac{1}{2}\beta_c w_{ic}^2\right) - \frac{N}{2}\log\beta_c + \kappa \tag{5.11}$$

$$-\log p(\mathbf{H}|\boldsymbol{\beta}) = -\sum_{c=1}^{C}\sum_{j=1}^{N}\log \mathcal{HN}(0,\beta_c^{-1}) = -\sum_{i=1}^{N}\sum_{c=1}^{C}\left(\frac{1}{2}\beta_c h_{cj}^2\right) - \frac{N}{2}\log\beta_c + \kappa \tag{5.12}$$

Therefore β_c manages the importance of community c in explaining the observed interactions. So, large values of β_c denote the column of **W** and row of **H** are close to zero and represent a community that does not explain the observed data. We assume that β_c are independent, and so place a standard Gamma distribution over them with fixed hyper-hyperparameters a, b [Penny and Roberts, 2002]. The negative log hyper-priors are:

$$\log p(\boldsymbol{\beta}) = \sum_{c=1}^{C}\log \mathcal{G}(\beta_c|, b) = \sum_{c=1}^{C}(\beta_c b - (a-1)\log\beta_c) + \kappa \tag{5.13}$$

Finally, the objective function $\mathcal{U}$ of Equation 5.6 can be expressed as the sum of Equation 5.10 through 5.13:

$$\mathcal{U} = \sum_{i}\sum_{j}\left[a_{ij}\log\left(\frac{a_{ij}}{\hat{a_{ij}}}\right)+\hat{a}_{ij}\right]+\frac{1}{2}\sum_{c}\left[\left(\sum_{i}\beta_c w_{ic}^2\right)+\left(\sum_{j}\beta_c h_{cj}^2\right)-2N\log\beta_c\right] + \sum_{c}(\beta_c b_c - (a_c-1)\log\beta_c) + \kappa \tag{5.14}$$

5.2.4 Parameter Inference

We now optimise Equation 5.14 to calculate **W**, **H** and $\boldsymbol{\beta}$. We follow [Psorakis et al., 2011, Tan and Fevotte, 2013, Lee and Seung, 1999, Berry et al., 2007], using the fast fixed-point algorithm presented in [Tan and Fevotte, 2013]. The algorithm updates **W**, **H** and $\boldsymbol{\beta}$ consecutively until a convergence measure is reached. The measure can be a maximum number of iterations or an allowance on the cost function. Algorithm 1 presents pseudocode of the fast fixed-point algorithm. The algorithm outputs $\mathbf{W}_\star \in \mathbb{R}^{N\times C}$ and $\mathbf{H}_\star \in \mathbb{R}^{C\times N}$, for which $\hat{\mathbf{A}} = \mathbf{W}_\star\mathbf{H}_\star$ represents the expectation network given the observed data and prior assumptions. The solution also consists of $C_\star$ the number of inferred communities.

Algorithm 1 Community Detection using NMF

```
Require: adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$, initial $C_0$, fixed Gamma hyperparameters a,b.
Define: matrix operation $\frac{\mathbf{X}}{\mathbf{Y}}$ as element-by-element division.
Define: matrix operation $\mathbf{X} \cdot \mathbf{Y}$ as element-by-element multiplication.
Define: $\mathbf{B} \in \mathbb{R}^{C \times C}$ as a matrix with elements $\beta_c$ in the diagonal and zero elsewhere.
 1: Auxiliary inputs $\mathbf{W}_0, \mathbf{H}_0$ from previous runs. If not present initialise to random values.
 2: for $i = 1$ to $n_{iters}$ do
 3:     $\mathbf{H} \leftarrow \left(\frac{\mathbf{H}}{\mathbf{W}^\mathsf{T}\mathbf{1}+\mathbf{B}\mathbf{H}}\right) \cdot \left[\mathbf{W}^\mathsf{T}(\frac{\mathbf{A}}{\mathbf{W}\mathbf{H}})\right]$
 4:     $\mathbf{W} \leftarrow \left(\frac{\mathbf{W}}{\mathbf{1}\mathbf{H}^\mathsf{T}+\mathbf{W}\mathbf{B}}\right) \cdot \left[(\frac{\mathbf{A}}{\mathbf{W}\mathbf{H}})\mathbf{H}^\mathsf{T}\right]$
 5:     $\beta_c \leftarrow \frac{N+a-1}{\frac{1}{2}\left(\sum_i w_{ic}^2+\sum_j h_{cj}^2\right)+b}$
 6: end for
 7: $C_\star \leftarrow \#$ of non-zero columns of $\mathbf{W}$ or rows of $\mathbf{H}$
 8: $\mathbf{W}_\star \leftarrow \mathbf{W}$ with zero columns removed
 9: $\mathbf{H}_\star \leftarrow \mathbf{H}$ with zero columns removed
10: return $\mathbf{W}_\star \in \mathbb{R}^{N \times C}$, $\mathbf{H}_\star \in \mathbb{R}^{C \times N}$
```

5.2.5 Complexity

The computational load of Algorithm 1 is mainly due to the matrix multiplication **WH**, which can be found in steps 3 and 4, and, therefore, the order is $\mathcal{O}(N^2C)$. However, we can exploit the sparse nature of real network adjacency matrices [Clauset et al., 2004]: the dot products $\sum_c w_{ic}h_{ci}$ where a_{ij} equals zero need not be calculated, hence reducing the complexity to N^2.

NMF operates on the entire adjacency matrix and so can be memory inefficient when implemented naively. The quadratic complexity can be reduced by loading only certain columns/rows of A into memory when needed, as Algorithm 1 performs no holistic operations for **A** or $\hat{\mathbf{A}}$ (such as inverse or multiplication) [Psorakis et al., 2011].

5.3 Applications

5.3.1 Directed Graphs

Following the discussion on NMF for directed graphs, in Subsection 5.2.1, we provide a simple example of how more information can be extracted in the directed setting versus the undirected setting. Consider the simple network in Figure 5.2, which consists of 8 nodes and 7 edges—if we ignore link directionality and run the NMF algorithm for community detection, we gain a partition on the dashed line where each community corresponds to a 'star'-like community.

We then run the algorithm with link directionality (asymmetric adjacency matrix) and discover three communities as shown in Figure 5.2. The community one consists of nodes 2, 3, 4 (blue), community two associates nodes 1, 6, 7, 8 (red), while community three consists only of node 5. In this setting, the communities are no longer comprised of mutually connected nodes but correspond to groups where members point to a common target [Psorakis et al., 2011].

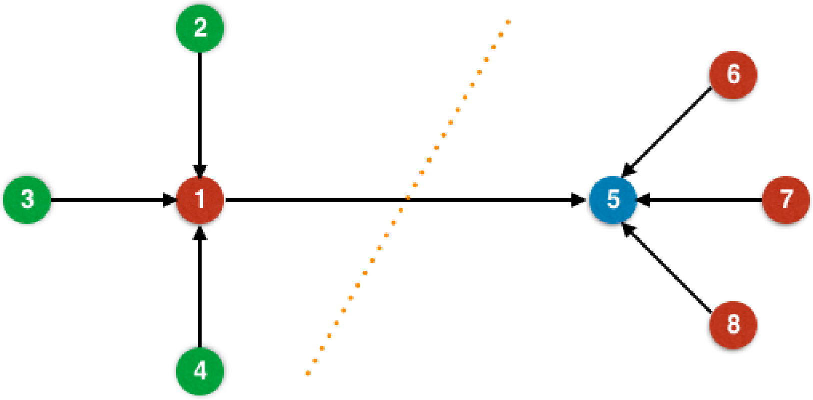


Figure 5.2: An example simple directed graph consisting of 8 nodes and 7 edges. When link directionality is ignored, two communities are extracted shown by the dashed orange line. In the directed case we extract three communities shown by the colours of the nodes. In this setting, the communities are no longer comprised of mutually connected nodes but correspond to groups where members point to a common target.

5.4 Discussion

In this chapter, we described a Bayesian non-negative matrix factorisation method for community detection, which was first presented by [Psorakis et al., 2011], in which community detection can be viewed as a generative model in a probabilistic framework. We also showed how this algorithm has a competitive advantage against other community detection methods, as it can be applied to directed networks, hence we are detecting information flow. The algorithm produces a soft-partition membership across communities, which is more reflective of real world problems. In Chapter 6, we apply the algorithm to the Enron Corpus.

Chapter 6

Analysis of the Enron Corpus

6.1 Introduction

In this chapter, the focus is on analysing the Enron Corpus from a social network perspective. The aim is to show that the techniques introduced in the previous chapters can be applied to communication networks to examine sociality on a community and individual level. The sliding window introduced in Chapter 4 is used to produce a stack of adjacency matrices. The latent communities are then extracted using NMF introduced in Chapter 5. We further analyse these communities using entropy and concordance, finally, we combine this with the net emails sent. Following this we explore how we can identify the most influential nodes as well as tracking nodes through time.

6.2 Graph Extraction

Before performing our analyses, we are required to extract the social graphs from the transactional data as presented in Section 4. The data consists of 151 employees and spans 2 years from January 2000 to December 2001.

Instead of applying our analysis directly to the entire dataset, we break the data into time windows of 103 days with the start dates incremented by one day, as described in Chapter 4, where we show that this provides the maximally informative network. For the entire dataset, the median number of emails per slice is 8285, the mean is 8602, and the standard deviation is 4394.

To complete, we construct directed weighted adjacency matrices, by adding 1 to position a_{ij}, every time node i sends an email to node j. Therefore turning our data from a transactional database into a stack of adjacency matrices $\{A^{(t)}\}_{t=1}^{T}$ that represent the Enron social network. Figure 6.1 shows an example of the Enron network on a given time period in 2001.

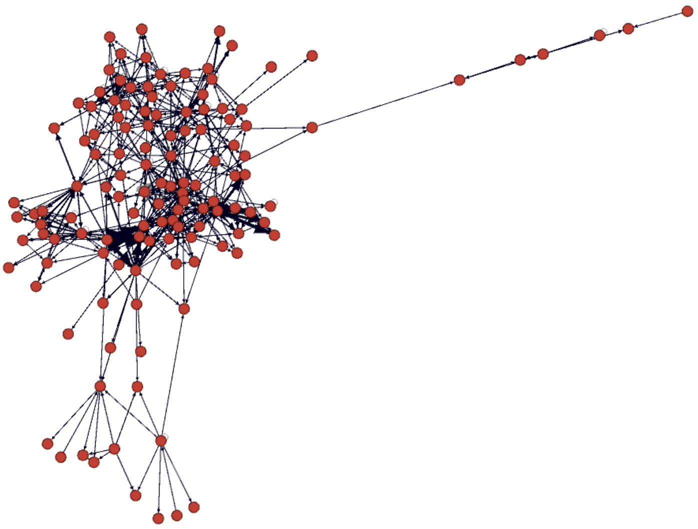
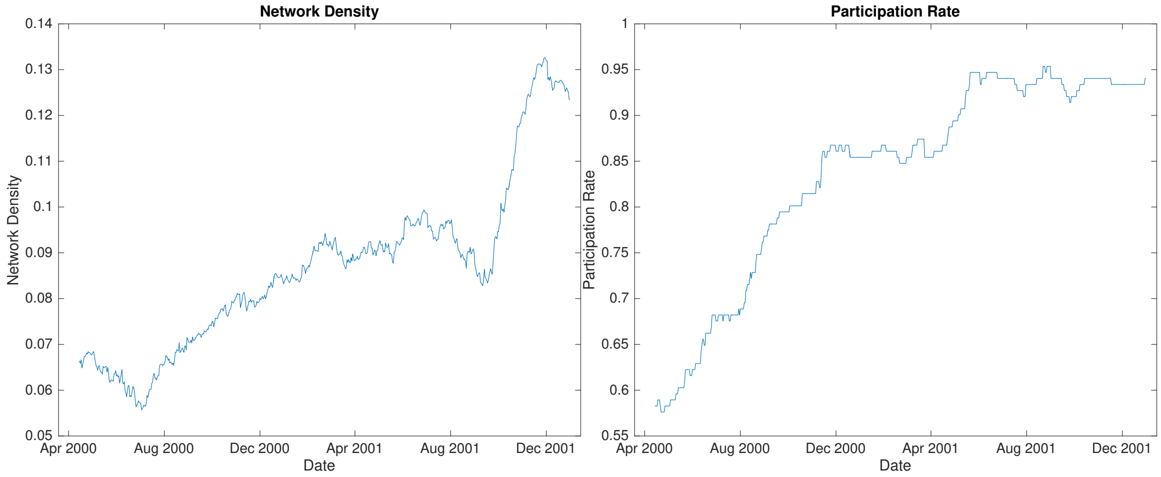


Figure 6.1: An example of the Enron network on a given time period in 2001

6.3 Analysis of Network Connectivity

We commence our analyses by exploring the global topological properties of our temporal network, and this starts by investigating Network Density (ND), which is the ratio of actual connections to the potential connections, $\frac{1}{2}n_t(n_t - 1)$. The stack of adjacency matrices are converted from weighted to unweighted before calculating the network density. We also record participation rate, defined as the fraction of population coverage of our data at time t, $PT = \frac{N^{(t)}}{N_{\text{total}}}$. Figure 6.2 presents the network density and participation rate of each adjacency matrix $A^{(t)}$ in the stack. The network density shows that we yield sparse adjacency matrices which is expected of a real-world network [Clauset et al., 2004]. The participation rate start at around 60% and reaches nearly 100% at the height of the crisis, and this will be important when analysing the latent communities.



(a) Network Density

(b) Participation Rate

Figure 6.2: We plot the network density and participation rate of each adjacency matrix.

6.4 Analysis of Community Structure

Having generated the "snapshot" adjacency matrices $\left\{A^{(t)}\right\}_{t=1}^{T}$, we ran the NMF community detection algorithm explained in Chapter 5. This output estimates for **W** and **H**, with these we greedily assigned each individual i to a community c_i to which they "most" belong, i.e. $c_i = \text{argmax}_{c \in c} w_{ic}$. Different community assignments were computed for some learners due to random initialisations of **W** and **H**, as well as numerical precision issues that affect the group allocation step. To reduce this we ran the algorithm 20 times. The time window size used was 103 days and moved through time with one day steps. In this section we will also explore the changes in the number of communities; how these communities are changing and how the "fuzziness" of the network changes.

6.4.1 Communities

We extract the communities via the use of CD-NMF by applying the algorithm on each time sliced adjacency matrix. Firstly we examine how the number of communities varies through time. It is important to note that the algorithm places all the nonparticipating nodes in a single community. In Figure 6.3, we can see that during 2000 the number communities stands around 15 (with 60% participation rate) and after the crisis communities dropped to 12 communities (near 100% participation rate), suggesting the network has become more centralised, communication and group cohesion has increased.

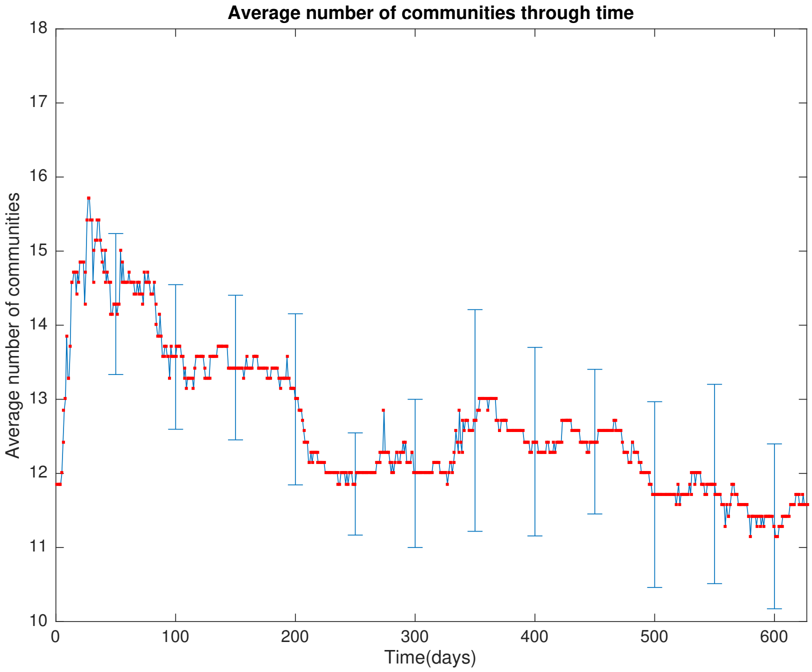
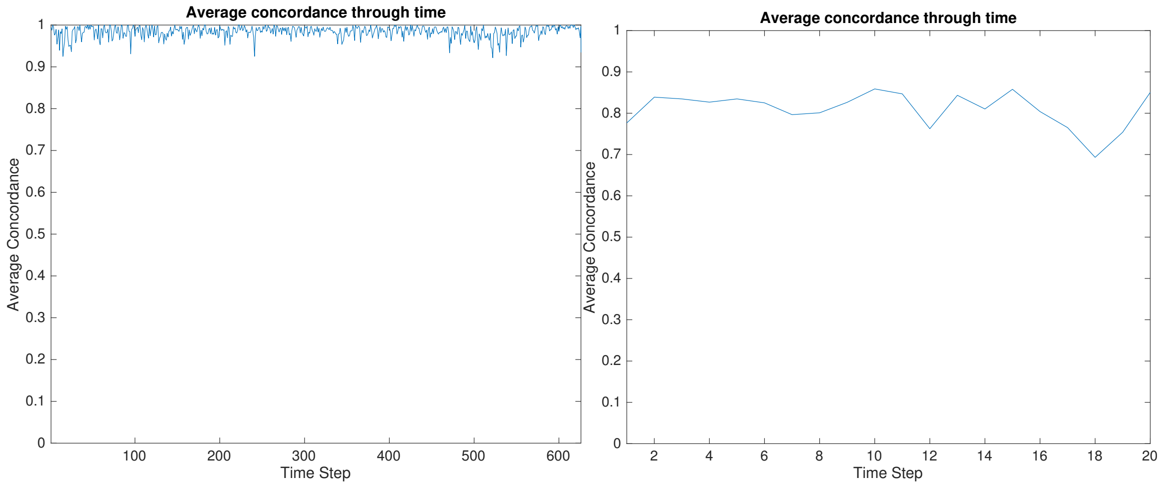


Figure 6.3: Shows the average number of communities through time extracted using a window size of 103 days. Every 50 days there is a vertical errorbars which shows the standard deviation.

6.4.2 Community Concordance

As we need to monitor how the membership of these communities are changing in the temporal setting, we undertake this with a concordance measure. The measure takes two arrays (communities) and then takes the ratio of the number of the same elements to the total number of elements. In this case, we are comparing the members of communities from one time step to the next. In Figure 6.4a we show the concordance for each time step of size 1 day across the entire dataset. It shows that the concordance between time steps stays above 0.9. This is expected as we are sliding the window one day at a time so the network will evolve slowly and be stable. In Figure 6.4b, we alter the time step to 30 days and so we are comparing the concordance of the communities month on month. As expected the concordance decreases to a value of approximately 0.8. The concordance values month on month are similar again suggesting the slow evolution of the network and the stability even during the crisis.



(a)

(b)

Figure 6.4: (a) the concordance measure with a time step of 1 day. (b) the concordance measure with a time step of 30 days.

6.4.3 Entropy of the Network

The next step is to investigate the "fuzziness" of the network. In the NMF framework, each node is given a membership distribution, and therefore, community allocation is not a binary decision but a belief. For example, mediator nodes, with high "betweeness", have a more entropic distribution. Having a soft-membership distribution allows us to assign how confident we are in assinging node i to community kc but also to quantify the "fuzziness" of the network. This is done by taking the average relative entropy of the node membership distributions. The relative entropy is defined as:

$$H = H_0 - H^{WH}, \tag{6.1}$$

where $H_0 = \log_2 C$ and C is the number of communities.

Figure 6.5 shows the plot of average entropy through time. At the beginning, the algorithm is very uncertain about the communities, and then through time the entropy drops and levels out between 2.5 and 3 bits. Furthermore, Figure 6.6 shows six snapshots of the communities and the strength of connection of the nodes within each community at different times during the lifecycle of the dataset. The figure shows that communities become crisper and the strength of ties within each community increases. Combining this with the decrease in entropy, it suggests that after the crisis there was an increase in group cohesion, suggesting that inequality of importance increased and managerial roles became more directive. This will be explored more in Subsection 6.5 where we explore the entropy values on a node level.

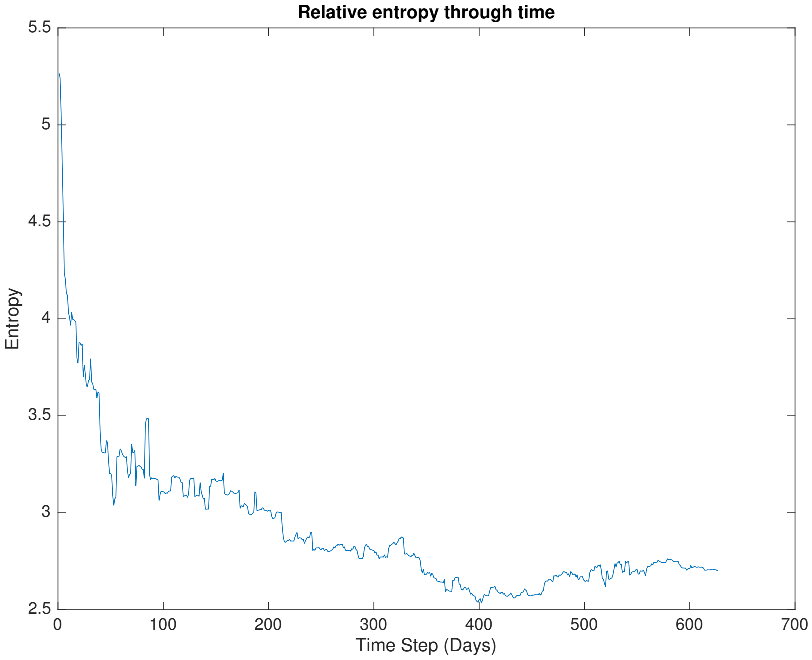
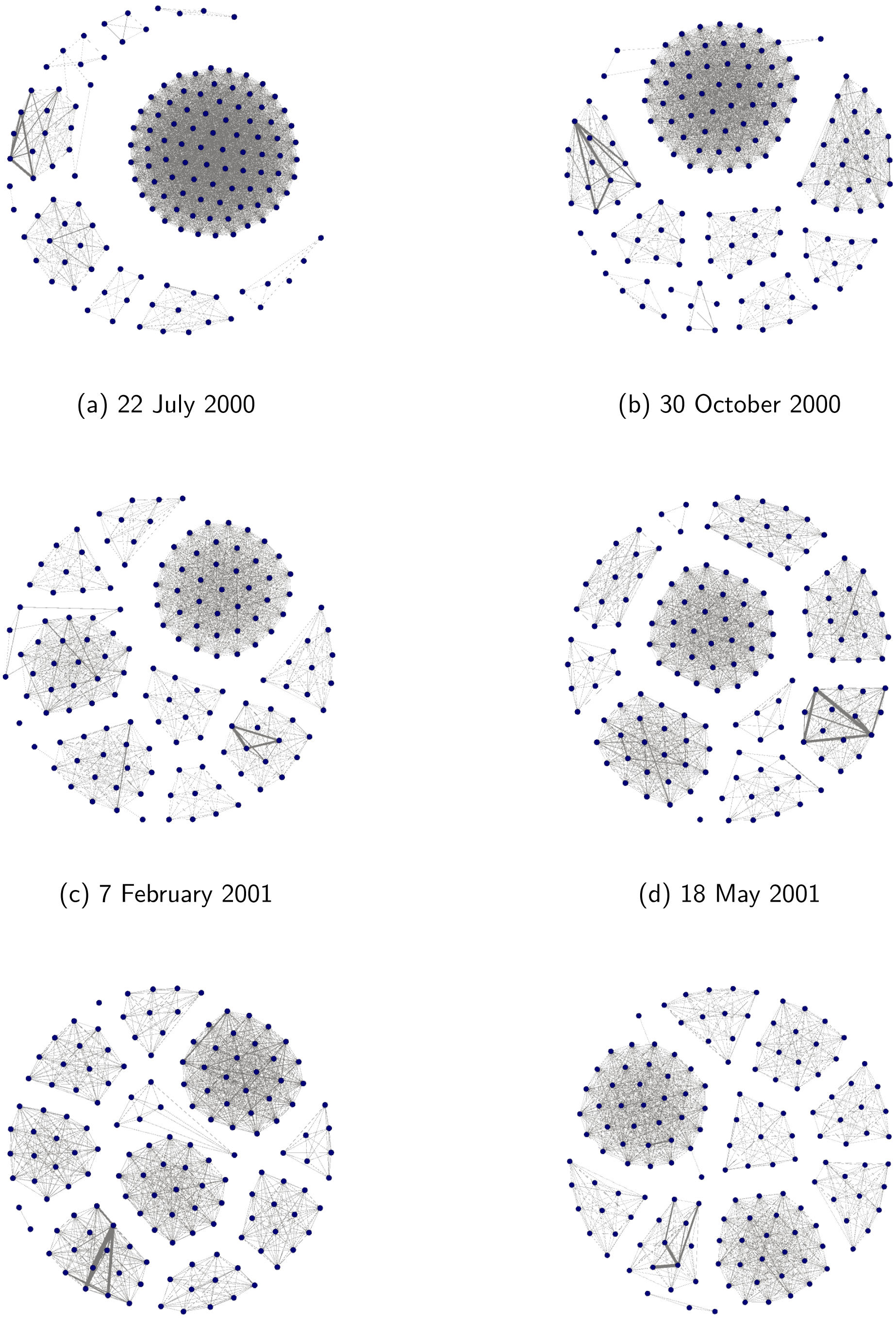


Figure 6.5: The relative entropy plotted through time

6.4.4 Net Senders vs. Receivers

Using a direct network framework allows us to track the number of emails that each community or node sends and receives, which enables us to distinguish whether a community or node is a source or sink of information. this provides a better understanding of information flow and combined with fact a community is a "mediator", we can gain a very good understanding of the most influential nodes within the network. Figure 6.7 shows an example snapshot of the net senders and net receivers (on a community level) and to whom they are communicating.



(a) 22 July 2000

(b) 30 October 2000

(c) 7 February 2001

(d) 18 May 2001

(e) 26 August 2001

(f) 4 December 2001

Figure 6.6: Snapshots of how the communities change through the lifecycle of the data set.

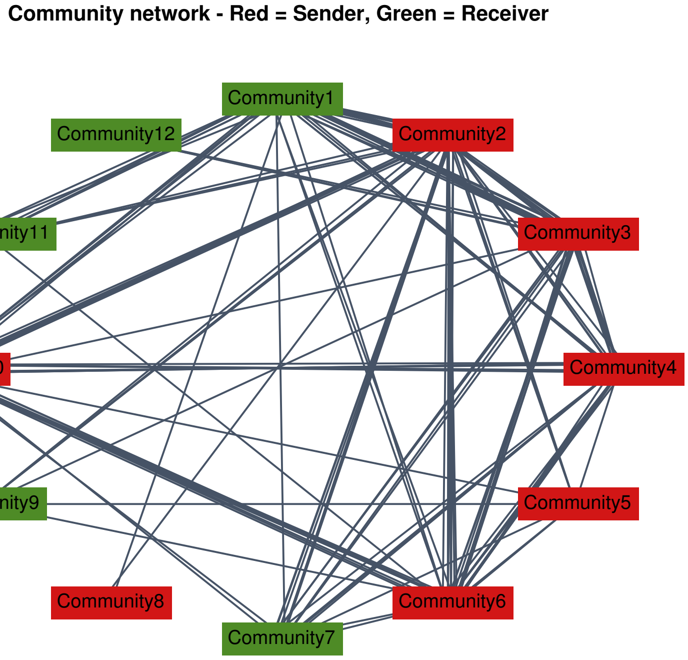


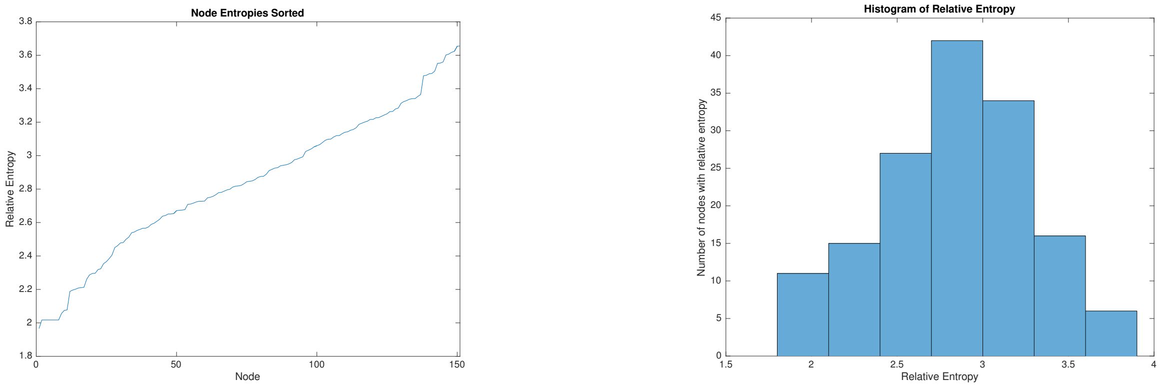
Figure 6.7: Shows an example snapshot of the net senders and receiver (on a community level) and to whom they are communicating

6.5 Tracking Individuals

In our framework, we convert the network from being node-to-node to being a node-to-community matrix. This allows us to track individuals much more easily by manipulating the soft-membership-matrix. In this section, we will track example individuals, in order to show how we can identify influential nodes. We will explore at a node level the membership, entropy, before finally providing case study individuals.

6.5.1 Entropy of Individuals

Subsection 6.4.3 discussed average entropy of the soft-membership matrix to quantify the "fuzziness" of the network. The entropy of the node-membership distribution can be used to identify "core" nodes in communities that act as mediators between different groups. The ability to identify core nodes in a directed network becomes more powerful when combined with net links (in this case emails), as we can discover the key influencers within the network. For example, in the Enron context it is reasonable to suggest that someone is an influencer if they have high entropy and are a net sender. The distribution of entropies by node and a histogram are show in Figure 6.8. This shows that the entropies vary from around 2 up to 3.6 and that these are distributed approximately normally.



(a) Node entropies sorted.

(b) Histogram of relative entropy.

Figure 6.8: Shows the relative entropies of the nodes sorted and a histogram of the relative entropies.

6.5.2 Concordance of Individuals

In order to track an individual's concordance score, we created a membership function. It finds the community of the node for two input time steps and calculates the concordance score. This measure provides us an understanding of whether the individual is staying part of one group or is moving groups. We use this to analyse example individuals in the next section.

6.5.3 Case Studies

A CEO

In December 2000, Jeff Skilling (COO) took over from Kenneth Lay as CEO. However, by August 2001, Skilling surprisingly resigned. The entropy plot starts at around 5 then drops down to a value of 2 bits as the algorithm becomes more certain about his position in the Company. However, his entropy then increases to just under 4 bits for 6 months during suggesting he has more influence within the network. This increase is the period in which he becomes CEO. The concordance plot for Jeff Skilling shows that he is moving communities often; again suggesting someone with a lot of influence within the network.

Louise Kitchen

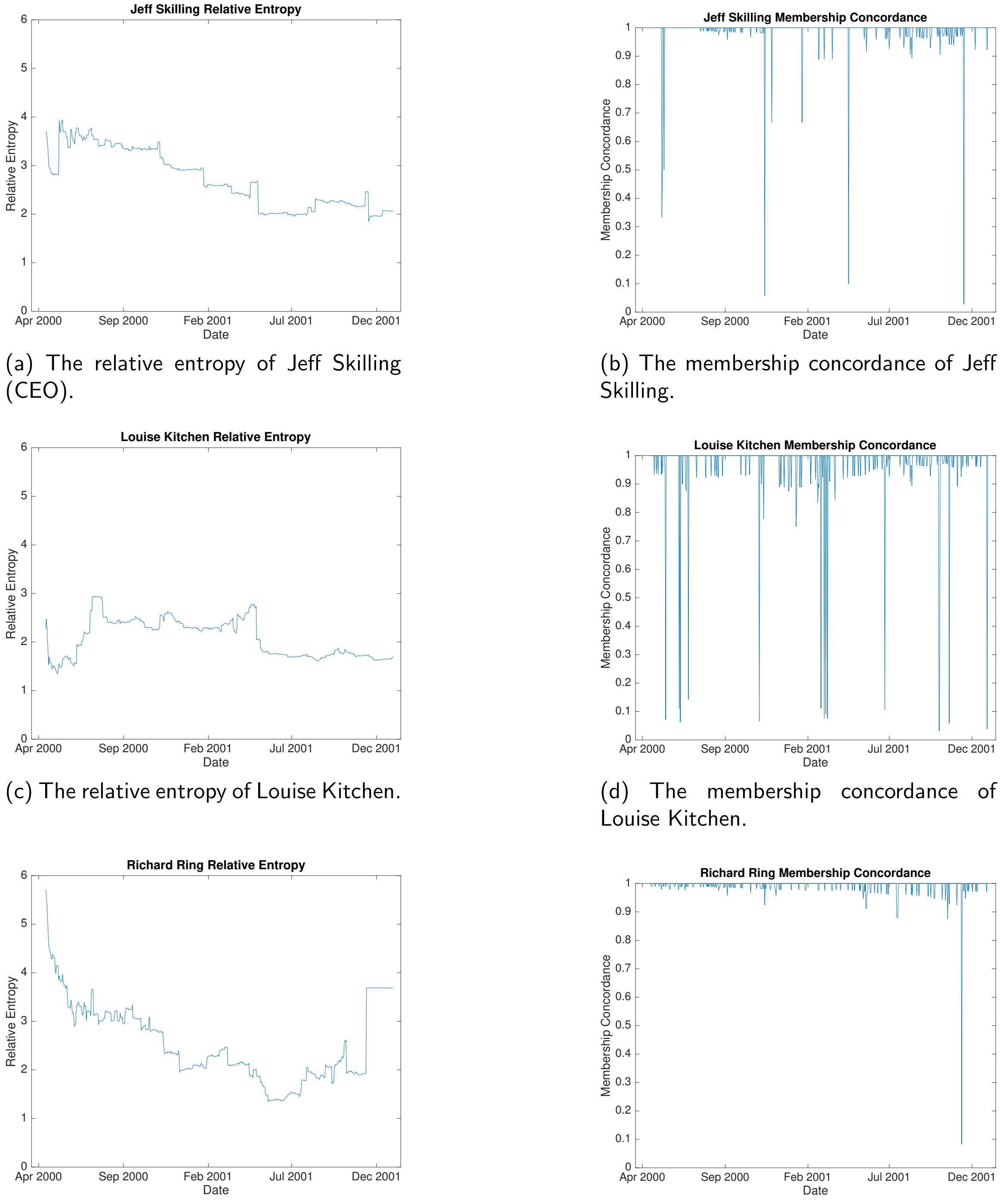
Louise Kitchen was a young British trader spearheading Enron's entry into Europe's energy markets. She was not a top executive and was under the age of 30 years. For someone with this profile we would expect them to not be that influential, therefore low entropy and high concordance. However, Kitchen rallied the support of Enrons best commercial, legal and technical people to work on companys online trading operation after normal working hours, and without the knowledge of any executives. What is interesting is her concordance plot shows that she keeps moving communities, and this suggests she is a mediator node and obviously in contact with many parts of the company.

An Employee

Finally, Richard Ring was just an employee; the algorithm suggests that he is part of one community for the entirety of the dataset and has an average entropy of 2.2 bits, and this is what would be expected of a low level employee.

6.6 Discussion

In this chapter, the methodological techniques introduced in this report were applied in order to explore the Enron Corpus. It was run on a dataset from January 2000 to December 2001 and extracted social networks for a time window of size 103 days. We followed this by extracting the communities using NMF — showing how it can be applied to a real directed dataset. As a final step, we investigated techniques for exploring these communities.



(a) The relative entropy of Jeff Skilling (CEO).

(b) The membership concordance of Jeff Skilling.

(c) The relative entropy of Louise Kitchen.

(d) The membership concordance of Louise Kitchen.

(e) The relative entropy of Richard Ring (employee).

(f) The membership concordance of Richard Ring.

Figure 6.9: The entropy and membership concordance plots of three example employees.

Chapter 7

Conclusions

Network analysis has been adopted as an excellent modelling framework for a wide variety of application domains, such as the Web, social networks and communication networks as it provides a range of methodologies for exploring the interconnections between individuals. Through network analysis, we have been able to explore the structural properties of the Enron Corpus. We have demonstrated how a time window approach can be used to infer the social structure of the dataset. Having constructed the network, we extracted the latent communities in a directed setting using Bayesian nonnegative matrix factorisation for community detection to explore the relationships between employees during the crisis. We then explored how these communities changed through time, using various measures such as entropy and concordance. The work in this report does not aim to be a theoretical contribution to the field of sociology. However, we investigated a series of technical questions, that arise from the communication datasets such as the Enron Corpus. In conclusion, this report accomplishes the general data analysis aims outlined at the beginning.

Bibliography

[Aggarwal and Yu, 2005] Aggarwal, C. C. and Yu, P. S. (2005). Online analysis of community evolution in data streams. *Proceedings of the 2005 SIAM International Conference on Data Mining*, page 5667.

[Albert et al., 1999] Albert, R., Jeong, H., and Barabsi, A.-L. (1999). Internet: Diameter of the world-wide web.

[Bastian et al., 2009] Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks.

[BBC, 2002] BBC (2002). Business — Enron scandal at-a-glance.

[Berry et al., 2007] Berry, M. W., Browne, M., Langville, A. N., Pauca, V. P., and Plemmons, R. J. (2007). Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics and Data Analysis*, 52(1):155173.

[Bifet and Gavald, 2007] Bifet, A. and Gavald, R. (2007). Learning from time-changing data with adaptive windowing. *Proceedings of the 2007 SIAM International Conference on Data Mining*, page 443448.

[Buchanan and Caldarelli, 2010] Buchanan, M. and Caldarelli, G. (2010). A networked world. *Physics World*, 23(02):2224.

[Calo, 2002] Calo (2002). Cognitive assistant that learns and organizes.

[Clauset et al., 2004] Clauset, A., Newman, M. E. J., and Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, 70(6).

[Cohen and Havlin, 2003] Cohen, R. and Havlin, S. (2003). Scale-free networks are ultrasmall. *Physical Review Letters*, 90(5).

[Diesner et al., 2005] Diesner, J., Frantz, T. L., and Carley, K. M. (2005). Communication networks from the enron email corpus it's always about the people. enron is no different. *Computational and Mathematical Organization Theory*, 11(3):201228.

[Granovetter, 1977] Granovetter, M. S. (1977). The strength of weak ties. *Social Networks*, page 347367.

[Hagberg et al., 2008] Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena, CA USA.

[Holme, 2003] Holme, P. (2003). Network dynamics of ongoing social relationships. *EPL (Europhysics Letters)*, 64(3):427.

[Klimt and Yang, 2004a] Klimt, B. and Yang, Y. (2004a). The enron corpus: A new dataset for email classification research. *Machine Learning: ECML 2004 Lecture Notes in Computer Science*, page 217226.

[Klimt and Yang, 2004b] Klimt, B. and Yang, Y. (2004b). Introducing the enron corpus. In *First Conference on Email and Anti-Spam (CEAS) Proceedings*.

[Krings et al., 2012] Krings, G., Karsai, M., Bernhardsson, S., Blondel, V. D., and Saramki, J. (2012). Effects of time window size and placement on the structure of an aggregated communication network. *EPJ Data Science*, 1(1).

[Lee and Seung, 1999] Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by nonnegative matrix factorization. *Nature*, 401:788–791.

[Mackay, 1995] Mackay, D. (1995). Probable networks and plausible predictions a review of practical bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, 6(3):469505.

[Milgram, 1967] Milgram, S. (1967). The small-world problem. *PsycEXTRA Dataset*.

[Nasraoui et al., 2003] Nasraoui, O., Uribe, C., Coronel, C., and Gonzalez, F. (2003). Tecno-streams: tracking evolving clusters in noisy data streams with a scalable immune system learning model. *Third IEEE International Conference on Data Mining*.

[Newman, 2003] Newman, M. (2003). The structure and function of complex networks.

[Newman, 2004] Newman, M. E. J. (2004). Analysis of weighted networks. *Physical Review E*, 70(5).

[Newman, 2015] Newman, M. E. J. (2015). *Networks an introduction*. Oxford Univ. Press.

[Newman and Girvan, 2004] Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69(2).

[Newman et al., 2001] Newman, M. E. J., Strogatz, S. H., and Watts, D. J. (2001). Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64(2).

[Penny and Roberts, 2002] Penny, W. and Roberts, S. (2002). Bayesian multivariate autoregressive models with structured priors. *IEE Proceedings - Vision, Image, and Signal Processing*, 149(1):33.

[Porter and Gleeson, 2016] Porter, M. A. and Gleeson, J. P. (2016). Dynamical systems on dynamical networks. *Frontiers in Applied Dynamical Systems: Reviews and Tutorials Dynamical Systems on Networks*, page 4951.

[Psorakis, 2013] Psorakis, I. (2013). *Probabilistic inference in ecological networks; graph discovery, community detection and modelling dynamic sociality*. PhD thesis, University of Oxford.

[Psorakis et al., 2011] Psorakis, I., Roberts, S., Ebden, M., and Sheldon, B. (2011). Overlapping community detection using bayesian non-negative matrix factorization. *Physical Review E*, 83(6).

[Qian et al., 2006] Qian, R., Zhang, W., and Yang, B. (2006). Detect community structure from the enron email corpus based on link mining. *Sixth International Conference on Intelligent Systems Design and Applications*.

[Rosvall, 2006] Rosvall, M. (2006). *Information horizons in a complex world*. PhD thesis, Doctoral thesis, Umeâ University, Faculty of Science and Technology, Physics, ISBN 9172641177.

[Sarkar and Moore, 2005] Sarkar, P. and Moore, A. W. (2005). Dynamic social network analysis using latent space models. *ACM SIGKDD Explorations Newsletter*, 7(2):3140.

[Shetty and Adibi, 2004] Shetty, J. and Adibi, J. (2004). The Enron email dataset database schema and brief statistical report. *Information Sciences Institute Technical Report, University of Southern California*.

[Spiliopoulou, 2011] Spiliopoulou, M. (2011). Evolution in social networks: A survey. *Social Network Data Analytics*, page 149175.

[Tan and Fevotte, 2013] Tan, V. Y. F. and Fevotte, C. (2013). Automatic relevance determination in nonnegative matrix factorization with the /spl beta/-divergence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):15921605.

[Watts and Strogatz, 1998] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of small-world networks.