Partitioning Massive Graphs for Content Oriented Social Network Analysis

Inaugural-Dissertation
zur
Erlangung des Doktorgrades der
Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von
Maximilian Viermetz
aus London, U.K.

09.07.2008

Aus dem Institut für Informatik
der Heinrich-Heine Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Referent:	Prof. Dr. Stefan Conrad Heinrich-Heine-Universität Düsseldorf, Germany
Koreferent:	Prof. Dr. Dietmar Seipel Julius-Maximilians-Universität Würzburg, Germany
Tag der mündlichen Prüfung:	12.06.2008

To Felicitas. for all she has given.
To Caroline, for suffering my involvement.
To Kurt, for his unconditional support.

And since you know you cannot see yourself,
so well as by reflection, I, your glass,
will modestly discover to yourself,
that of yourself which you yet know not of.

— *William Shakespeare*
Julius Caesar
Act I Scene II

Zusammenfassung

Heutzutage sind Firmen in einer immer enger vernetzten Welt einer stetig wachsenden Informationsflut ausgesetzt. Berichte, Nachrichten und Konsumentenkommentare sind jederzeit und überall über Plattformen wie Weblogs oder Newsgroups verfügbar. Die Möglichkeit der Kommunikation über elektronische Medien wird immer stärker genutzt. Diese gewinnen dadurch immer mehr Einfluss. Preisvergleiche zwischen verschiedenen Anbietern, Testberichte von professionellen Agenturen sowie der Austausch von Benutzern über gemeinsame Erfahrungen und Beschwerden sind nur einige Bereiche in denen moderne Kommikationsmedien eine immer stärkere Rolle spielen.

Eine Konsequenz der rasanten Entwicklung von modernen internetbasierten Kommunikationsmedien ist die starke Entwicklung und Nutzung von asynchronen Kommunikationsmodi, wie zum Beispiel Email oder Newsgroups, und synchronen Kommunikationsmodi wie sie durch Instant Messaging angeboten werden. Durch ihre bequeme Handhabung, eine so gut wie kostenlose Nutzung und eine nahezu sofortige Kommunikation haben sie die moderne Geschäftswelt komplett durchdrungen.

Die Bereitstellung moderner Kommunikationsinfrastruktur auf elektronischer Basis ermöglicht die Erfassung und Sammlung neuer Daten in ungeahntem Maße. Diese Datenquellen sind das primäre Ziel moderner Techniken im Bereich Data Mining und Informationsextraktion. Daraus entstehende Kommunikationskorpora werden seit einiger Zeit analysiert und unter dem Aspekt des Social Network Analysis eingehend betrachtet. Die Schwerpunkte dieser Betrachtung kann man in zwei Kategorien einteilen; erstens die Untersuchung der Struktur von Kommunikationsgraphen und zweitens die inhaltliche Analyse der Kommunikationen um vorher unbekannte aber jedoch interessante Informationen zu extrahieren.

Social Network Analysis findet in der heutigen Informatik immer mehr Anklang. Durch die Analyse von Kooperationsnetzwerken durch Publikationen (co-citation networks) wurde seit 1990 die Anwendung von Graphanalyse auf Netzwerke aller Art in Augenschein genommen.

Die Entdeckung und das Herausarbeiten neuer und markanter Trends und Marktentwicklungen können das Aufkommen von einflussreichen Themen und Schwerpunkten in Planung und Kommunikation von Unternehmen beeinflussen. Von besonderem Interesse ist vor allem die Beobachtung der Entwicklung solch relevanter Themen über die Zeit.

Die Resultate von Social Network Analysis werden heutzutage in den unterschiedlichsten Bereichen eingesetzt. So werden Netzwerkanalysen bei der Erkennung und Bekämpfung von terroristischen Netzwerken bis hin zur SNA basierten Bewertung und Sortierung des Internets durch Google, der besten Geschäftsidee des letzten Jahrhunderts, genutzt. Zudem wurden im Fachgebiet der Informationsverarbeitung in den letzten Jahren große

Fortschritte erzielt. Diese ermöglichen eine wesentlich präzisere automatische Analyse und Klassifikation von Texten.

Ansatz dieser Dissertation ist es grundlegende Aspekte aus Social Network Aanalysis und Informationsbverarbeitung zu vereinen und Kommunikationskorpora unter strukturellen sowie inhaltlichen Blickwinkeln gleichzeitig zu betrachten. Diese Arbeit konzentriert sich auf die Herausarbeitung von Kommunikationsmustern, welche sich aus dem Zusammenspiel von SNA und inhaltlicher Segmentierung durch den Austausch elektronischer Information ergeben. Der Fokus liegt primär in der Analyse von textbasierten Nachrichten, wird aber darüber hinaus weitere Quellen von gerichteten und textbasierten Kommunikationsformen betrachten.

Abstract

For companies acting on a global scale, the necessity to monitor and analyze news channels and consumer-generated media on the Web, such as web-logs and newsgroups, is steadily increasing. In particular the identification of novel trends and upcoming issues, as well as their dynamic evolution over time, is of utter importance to corporate communications and market analysts.

The development of the communication infrastructure as provided by electronic means and delivered via the Internet has provided increasing numbers of large data sources as a basis for analysis with the help of techniques provided by data mining and information retrieval. But not only has the number of data sources multiplied, so has the size and the complexity.

A particular pertinent aspect of the explosive growth of the Internet is the increasing utility and dependence upon electronic means of communication. The ease of use, speed and reliability has been quietly revolutionizing the way people conduct their day to day business since the inception and application in the latter part of the previous century.

Communication corpora have been considered for some time now. The focus of analysis has mainly fallen into two camps, the first looking at the structure of the communication graph, and the second using the content as a basis for information retrieval techniques to mine previously unknown facets from the data.

It is our opinion that the combination of content with structure can significantly increase the quality of data extracted as well as increase the flexibility when considering very large text corpora.

This thesis will focus on the analysis and use of the communication patterns which can be discovered in the flow of conversations over such electronic media. The focus will reside primarily on e-mail driven communication, but will be generalized to any form of communication fitting into description of directed and text-based communication. This allows a broad spectrum of communication methods to be considered.

Contents

1 Introduction

Communication via electronic media is almost ubiquitous today. The use of e-mail or instant messaging services has moved from being a feature to a necessity. This form of communication is of course amenable to the application of data mining techniques to extract and analyze communications content.

Communication networks today have achieved a global dimension, and the heavy use generates reams of data potentially to be scrutinized. This thesis is concerned with the understanding of large networks, such as can be found in corporations or large communication groups. Combined with time frames of several years, this implies the analysis of massive graphs, graphs of great size and complexity.

The work presented in this thesis lies at the intersection of several fields of research. On the one hand the use of standard methods developed to get a grip on the content of text are used as a basis for semantic analysis, whereas on the other hand the inclusion of social network analysis is used to apprise the structure of communication graphs. Upon these two pillars lies the work presented in this thesis.

Starting Point

The use of social network analysis has found increasing acceptance in the field of computer science. After the analysis of publication networks starting in mid 1990s by groups of physicists the application to other graph structures found in computer science has lead to somewhat of a boom in social network analysis. Applications range from terrorist network detection to the basic idea behind Google.

The use of natural language processing techniques to analyze text corpora has also made great strides in recent years. The aim has been to develop better and better techniques to automatically process text based data in order to perform data mining tasks such as topic discovery or content classification.

This work will use both avenues of exploration to tackle a specific problem encountered in modern text corpora: The size of communication corpora keeps growing exponentially.

Contribution

Text based messaging corpora have two distinct characteristics open to exploitation by automated exploration techniques. On the one hand specific statements about the structure of the communication can explicitly be derived from the communications graph. On the other hand content can be analyzed to find topics and other semantic aspects.

We propose a way to segment the data into useful slices. This thesis will use the semantic side to prise apart a communications corpus into several distinct sub-corpora. Each sub-corpus can then be analyzed with more detail and accuracy than the whole. We are able to perform semantic analyses based on content of the messages, and tie the content to individual topic based sub-networks of the overall communications network. This **topic based segmentation** of a communications network yields a more differentiated understanding of relationships within the network than a more general analysis of the entire network.

Thesis Structure

For this reason the introduction of related, and fundamental, work related to this thesis takes up most of the first half of this thesis. The background is introduced in chapter 2, in which the various sections touch the aspects of graphs (section 2.1), communication networks (section 2.2), as well as text mining (section 2.3) and related graph mining (section 2.4).

Chapter 3 proposes the central approaches of this thesis. Particularly the treatment of massive corpora and the interpretation of social network analyses based on such corpora are treated in sections 3.2 and 3.3 respectively. The introduction of methods geared toward the analysis of massive corpora is treated in sections 3.4 through 3.6. Chapter 4 examines case studies of the individual approaches from chapter 3, and chapter 5 examines the implementations and challenges presented by real world data in the context of this thesis.

2 Related Work

Contents

This thesis touches upon several disparate fields of computer science, on the one hand using graph theory as well as incorporating network analysis with the field of text analysis and text mining. The related work pertaining to this thesis therefore needs to be presented in the context of the approach proposed in chapter 3.3 on page 36. As this presents the background against which the thesis is set, this chapter can be referenced when necessary.

2.1 Graphs

Social Network Analysis is a field of research derived from work originally done in the social sciences. The further evolution and incorporation into computer science draws heavily on the fields of graph analysis as well as linear algebra for computational aspects. For this reason a specific introduction and definition of usage for each applicable field will form the background of this thesis.

Synchronous as well as asynchronous communication structure can be captured in graphs and consequently in adjacency matrices. This representation lies to the forefront of any analysis, such as either social network analysis or sub-graph detection. As graphs will play a fundamental and recurring role throughout this thesis an introduction of the material used is imperative.

The methodologies introduced here attempt to present a complete picture in terms of notation and concepts used, and is considered central to the understanding of the thrust of the approach presented in this thesis. It is suggested to the reader that the chapter can be referenced when needed.

2.1.1 Definitions and Usage

The understanding of social network analysis rests on the comprehensive work done in understanding graphs which form the basic structure capturing communications networks. For this purpose we first mention the definition of a graph as used in this context. While the introduction of graphs may seem tedious and unnecessary, they form a notational basis for the entire following work. For this reason we will start at the very beginning and introduce the elements of a communications network, the entities performing the communications and the ties between them.

Definitions

Usage

A number of definitions and common usages are mentioned in this section; whenever pertinent terminology is introduced it is clearly marked as such in the margin. First off the basic structure of graphs will be mentioned. The proposed notation underlies any discussion about the nature of graph based data. The concepts and definitions used in this thesis largely follow the ideas as described in [14].The following pages present a concise view of the basic definitions of graphs; this is not meant to be a complete reference, but is intended to lay the foundation for all future discussions.

Node

The primary entities of a graph are captured in a finite set of nodes. A node N is an entity in the context of a graph.

Edge

With the entities underlying a graph given as a group of nodes, the relationship between entities is described by associating them pairwise via a common connection. Given a set of nodes $\mathcal{N}$ where $n_1 \in \mathcal{N}$ and $n_2 \in \mathcal{N}$ are two nodes; the tuple $E = \{n_1, n_2\}$ represents a connection traversable in both directions between them. To distinguish the needed cases when a connection is traversable in one direction only, we denote the ordered pair $\overrightarrow{E} = (n_1, n_2)$ as a directed edge (in contrast to an undirected edge), connecting n_1 to n_2. An edge can carry a weight, indicating strength or importance depending on context. The weight $w \in \mathbb{R}^+$ of an edge $weight(E) = w$ is considered to be normalized to $w = 1$ in unweighted graphs. Weighting will play an important role during network analysis.

Weight

We can now work with entities as well as with relationships between them. Generally speaking we can now start to work with a graph as defined by a set of nodes $\mathcal{N}$ and an associated set of edges $\mathcal{E}$ over $\mathcal{N}$. A graph G is defined as the ordered set $G = (\mathcal{N}, \mathcal{E})$. In the context of this thesis we will be looking at finite graphs in which the set of nodes is finite as well as the set of edges.

Graph

Usage

Graphs will be considered extensively during the course of this work. In most cases during the implementation and analysis the representation chosen will not be the visual means such as in Figure 2.1(a) on page 7, but in a computational-friendly form of matrices. For completeness sake we make note of the equivalence between networks and the representation using an adjacency matrix, since any finite graph can be described with the help of an adjacency matrix. Given a finite graph $G = (\mathcal{N}, \mathcal{E})$ we can describe this network with an $n \times n$ matrix M where every entry $M_{ij} \geq 0$ describes whether there exists an edge between the nodes i and j. The weight $M_{ij} = w$ of an edge is taken to be 1 for unweighted graphs. A weight of $M_{ij} = 0$ indicates there exists no edge from node i to node j.

Adjacency Matrix

The representations of graphs used during the computational aspect of this thesis will revolve around adjacency matrices.

Graph Properties

Of interest are a few properties of a graph as well as of nodes within a graph. These aspects of graphs will be relevant in later stages of social network analysis, as well as pertaining to the related analytical tools.

The size of the data contained in graphs has a significant impact on the processing efficiency and effectiveness. Let $G = (\mathcal{N}, \mathcal{E})$ be a graph. For this reason we note the size of a graph $size(G)$ is then defined as $size(G) = |\mathcal{N}|$.

Graph Size

While the consideration of the number of nodes contained in a graph can characterize the size, we can also differentiate various classes of edge densities within a graph. This concerns the number of edges found in a specific specimen as compared to a graph of the same size with an edge between every pair of nodes. We designate two important cases concerning the later analysis, namely a graph $G = (\mathcal{N}, \mathcal{E})$ is considered a maximal graph G^* iff $\forall n_1, n_2 \in \mathcal{N}, n_1 \neq n_2 : \exists E(n_1, n_2) \in \mathcal{E}$

Maximal Graph

Likewise the density of a graph is taken to differentiate between strongly and sparsely connected graphs. The differentiation is a more fuzzy one, as the determination relies on the judgment of the number of edges found in relation to the maximum possible. Gen-

Sparse Graph

Dense Graph

erally a graph with *most* or on the other hand conversely *few* connections is considered to be of special interest to the analysis. Let G be a graph $G = (\mathcal{N}, \mathcal{E})$, and $G^*(\mathcal{N}, \mathcal{E}^*)$ be the associated maximal graph. G and is considered sparse iff $|\mathcal{E}| \ll |\mathcal{E}^*|$. In analogy a graph G is considered strongly connected when $|\mathcal{E}| \approx |\mathcal{E}^*|$.

Examples for sparse and maximal graphs can be seen in figures 2.1(a) and 2.1(b) on the facing page respectively.

Path

While the size of the data set to be analyzed is of considerable impact when efficiency is concerned, the circumference of a graph can provide an important indication as to the complexity of a graph at hand. Before we look at the circumference, we must first traverse between nodes of a graph. Let $n_1 \in G$ and $n_2 \in G$ be two nodes within a graph $G = (\mathcal{N}, \mathcal{E})$, and $\mathcal{M} = (m_1, \ldots, m_n) \subseteq \mathcal{N}$ an ordered set of nodes. Then a path between these nodes is considered to be $P(n_1, n_2) = (m_1, \ldots, m_n)$ iff

- $n_1 = m_1 \wedge n_2 = m_n$
- $\forall\ 1 \leq i \leq n-1 : \exists E = (m_i, m_{i+1}) \in \mathcal{E}$

Now that we can travel between nodes via edges, it is of interest to take note of a special case when considering these paths. By finding the shortest path between two nodes, we find paths analogous to the latitude between two points on a globe at the same height.

Geodesics

Let $n_1, n_2 \in \mathcal{N}$ be two nodes in a graph $G = (\mathcal{N}, \mathcal{E})$, and $\mathcal{P}$ the set of all paths between these two nodes. A subset of shortest paths $\overline{\mathcal{P}} \subseteq \mathcal{P}$ is considered to be $\overline{\mathcal{P}} = \{P \in \mathcal{P} | \forall Q \in \mathcal{P} : |P| \leq |Q|\}$ since there can be any number of shortest paths, any path $P_G \in \overline{\mathcal{P}}$ describes a geodesic for graph G.

It can be of interest to get an idea of how large and well connected a graph really is. To this end we can use the circumference of a graph to determine how long one needs to traverse the graph from one end to the other. The effective size of a graph is characterized by the average number of nodes to be traversed when traveling between any two given nodes. This has lead to the measuring of the World Wide Web, which as a graph [81] has a diameter of about 17 [5]. This means that barring unreachability the maximum hops one has to perform between any two web pages is not that large.

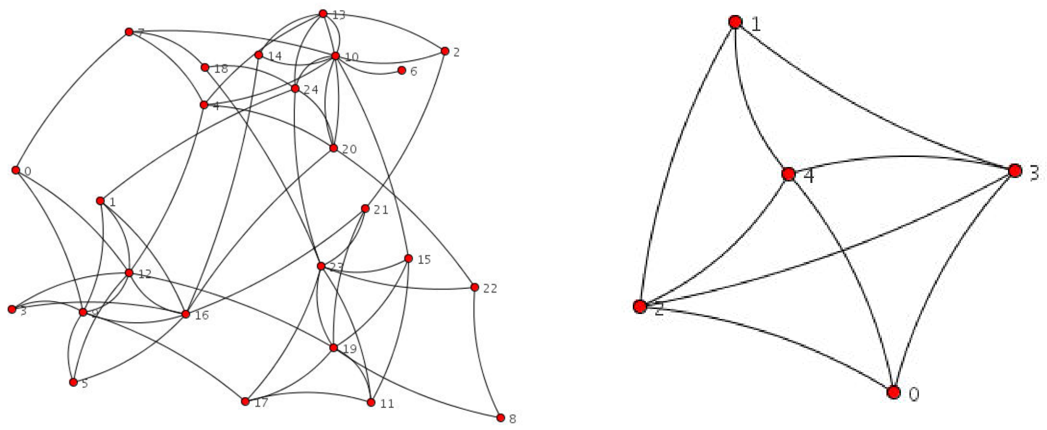
Graph Diameter

Using such shortest paths we can now classify the circumference of a graph. Let $\overline{\mathcal{P}}^*$ be the set of all shortest paths for all pairwise different nodes $m, n \in \mathcal{N}$ of graph $G = (\mathcal{N}, \mathcal{E})$. The set of longest direct paths $\mathcal{P}^{MAX} \subseteq \overline{\mathcal{P}}^*$ is considered to be $\mathcal{P}^{MAX} = \{P \in \overline{\mathcal{P}}^* | \forall Q \in \overline{\mathcal{P}}^* : |Q| \leq |P|\}$. The diameter $diam(G)$ is now the length of any longest direct path $P^{MAX} \in \mathcal{P}^{MAX}$: $diam(G) = |P^{MAX}|$.

Another recurring aspect of graphs is the notional neighborhood of a node. Especially in the context of social network analysis as well as graph mining we must often consider the nodes reachable from a given node, or the transitive case of reachable nodes in a given number of steps.

Node Neighborhood

In this thesis neighborhoods of varying sizes are of interest, these neighborhoods are centered on $n \in \mathcal{N}$, a node in a graph $G = (\mathcal{N}, \mathcal{E})$. A neighborhood of diameter $k \in \mathbb{N}$ is a subset $\mathcal{M} \subset \mathcal{N}$ where $\mathcal{M} = \{\forall m \in \mathcal{N} : \exists P(m, n) \wedge |P(m, n)| \leq k\}$ and is denoted by $neighborhood(k, n) = \mathcal{M}$.



(a) Sparse Graph

(b) Maximal Graph

Figure 2.1: Graphs

2.1.2 Graph Types of Interest

As we will be considering various communication networks, a basic differentiation between various graph types has been followed in this thesis. As we are focusing on network graphs generated by communications behavior, we will look at aspects concerned with exploring random and massive graphs.

Random Graphs

Introduced by Erdös and Renyi in the late 60s [39, 40], the basic idea is to create a graph from a given set of nodes by randomly adding edges to an at first empty graph. While the idea is simple, the patterns and properties described by this proposition have profound similarities to natural processes.

The perhaps most important aspect in the context of the present work is twofold [24]:

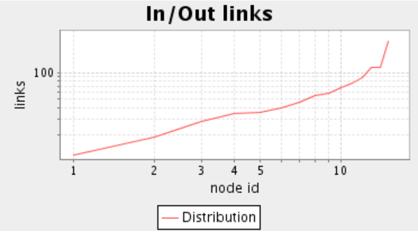
- The power-law distribution of links.
- The indication of sparse graph characterization.

The power law distribution of links has been observed in a wide variety of natural phenomena, and most importantly in this context in the distribution of links in the World Wide Web. Similarly, social networks display power law distributions among examples of collaboration and e-mail networks.

Scale Free Networks

Social networks display a distribution of in and outgoing links following a power law. These graphs are termed scale free as any random sub-graph will display a similar distribution of links per node as the overall graph. When viewing such a graph on any scale there will be few very densely connected nodes whereas the great majority are very poorly connected nodes. This observation impacts on the development of synthetic data. In chapter 5.3.3 we will use this property in constructing synthetic data for test and control purposes.

The adjoining figure shows the logarithmic distribution of links within a blogging network. The network itself is the example mentioned in 2.1(a), but examples of real world networks can be found in section 4 on page 53. We can see the quality espoused by scale-free networks, as there are significantly fewer nodes with many connections than with a small handful of connections[2]. This hub structure is discussed in detail by [24].



A common example of a scale free network is the link structure of the World Wide Web. Among the others mentioned by [12] is the very unusual electricity distribution network in the United States. As we know this fact became crucial in the cascade failure occurring in late 2003.

Sparsity

The power law distribution also implies a sparse graph for social networks, with only small subsets being complete graphs, or cliques. The implication of large sparse graphs is of interest to any calculation performed on the adjacency matrices during this thesis. Chapter 5.3.2 on page 92 shows the sizable problems posed by such large sparse matrices in a computational context.

Sub-graphs

Sub-graph

There are a few aspects of sub-graphs to network graphs that bear mentioning. Most aspects relating to social network analysis can be found in chapter 2.2.2 on page 18. These considerations are motivated by the need to find sub-graphs of denser consistency than the surrounding graph at large. Especially in large and massive graphs finding relevant sub-graphs becomes crucial to the interpretation of the analysis as well as the complexity of the problem analysis. Let $G = (\mathcal{N}, \mathcal{E})$ be a graph, then a sub-graph G' is considered to be $G' = (\mathcal{N}', \mathcal{E}')$ where $\mathcal{E}' \subset \mathcal{E}, \mathcal{N}' \subset \mathcal{N}$.

Maximal Sub-graph

There is one more variant to declare for notational purposes, the complete sub-graph. Let $G = (\mathcal{N}, \mathcal{E})$ be a graph, and $G' = (\mathcal{N}', \mathcal{E}')$ a sub-graph thereof. G' is considered a maximal sub-graph G'^{*} iff G' is a maximal graph.

This has implications upon the detection and classification of cliques, a central aspect to social network analysis. Chapter 3.3 on page 36 is concerned with finding sub-graphs which are not complete but can still be considered to describe cliques.

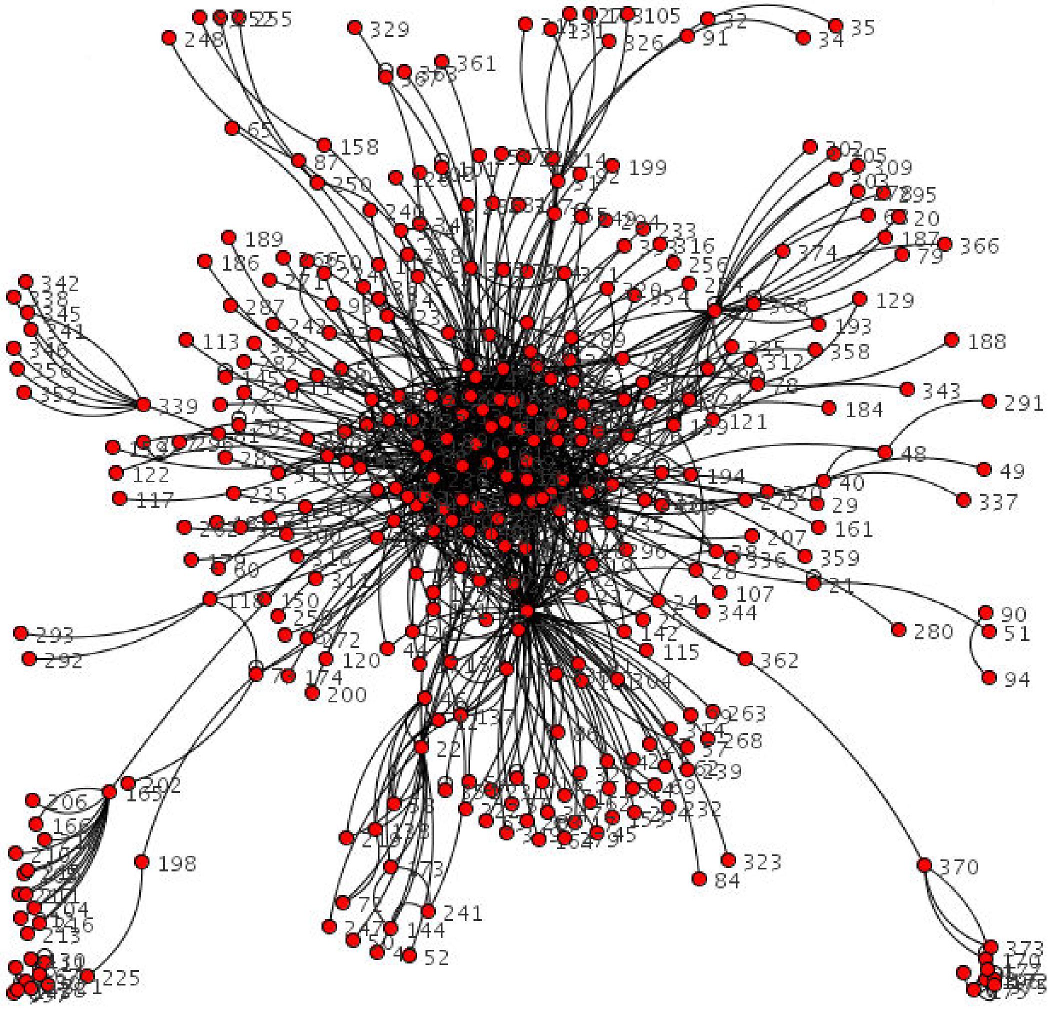


Figure 2.2: Massive Graphs: A Blogging Network

Massive Graphs

A central aspect of this work will be the treatment of massive graphs. In this section we will introduce the notion and related definitions of what will be considered a massive graph.

The workability of social network analysis in conjunction with content analysis indicates the use of the tools depend on the efficiency of both component fields. In this case the content analysis as envisioned and used can quickly become unwieldy and necessitate offline analysis.

It is one of our goals to impose constraints upon the data to facilitate the analysis in a timely fashion. For this reason we define massive graphs as graphs of sizes precluding the direct and timely analysis. At the time of writing a number of examples could be found:

Massive Graph

- The Internet as a Routing Network
- The World Wide Web
- Large Email networks
- Protein interaction networks

We specifically concentrate on email networks, as we will use the segmentation process to extract compact networks from massive graphs. Beyond the focus on email networks other networks based on instant messaging or other technologies generating a text-based communications corpus can all be taken into account by the approach engendered in this thesis.

Coherent Semantic Network

Any such corpus will be divided into meaningful sub-networks of similar content. These networks will be termed coherent semantic networks: Let $G = (\mathcal{N}, \mathcal{E})$ be a graph, and $G' = (\mathcal{N}', \mathcal{E}')$ a sub-graph thereof. We consider G' to be a **coherent semantic network** iff

- The content of G' is tightly defined in a topic.
- The content found in graph at large $G \setminus G'$ is dissimilar to this topic of G'.

2.2 Network Analysis

Social Network Analysis has provided a powerful and in recent years increasingly used set of tools useful when analyzing communication networks. The foundations and use of this field will be presented in this section.

2.2.1 Social Networks

A social network is based on a graph. Originally intended to capture the communication patterns of small groups, this approach can be generalized to accept any entity entering into a prescribed relation to other entities. This definition does not limit the tools of social network analysis to be applied to the analysis of actual people, but networks as a whole [124, 125, 97]; this thesis will follow the terminology used in the field of social network analysis.

The purpose of social network analysis is to bring actors into focus who have an extraordinary impact or importance within a communications network. For instance [46] talks about influence networks, specifically how the spread of information is driven by the relative persuasiveness of someone within a group.

In this context the terminology is subtly different from the context of graphs and graph analysis. The salient terms will be introduced on the following pages.

Definitions

While there is no practical difference to graph analysis, it has become a custom to differentiate between graphs and social networks through the use of different terms. This aids in the later interpretation of the graphs, as communications graphs are necessarily more interpreted than analyzed. Because we are talking about people, or more generally speaking, communicating agents, in the context of Social Network Analysis an entity partaking in the observed group is termed an **Actor**. A set of such actors $\mathcal{A}$ is the equivalent to a set of nodes in graphs. Actor

Instead of having edges between actors, the connection is generally considered to be a *link* denoting an observable exchange or influence between actors. In analogy to an edge in graphs, a connection between actors is termed a **Link**. A set $\mathcal{L}$ of links is the analog to the set of edges in graphs. Link

A group of actors and their links are considered to form a social network $\mathcal{SN} = (\mathcal{A}, \mathcal{L})$. Social Network

The combination of these components can now be said to yield a social network. This thesis does not necessarily equate a person with an actor, as this would limit the view unnecessarily. More generally any communicating entity can be considered an actor, so that not even a human being must take part in such a process. An actor can represent any party to the communication process.

Likewise the definition of a link does not have to be the sending of an email or equivalent exchange, nor does it have to be defined exchange of information. While this thesis looks at communication corpora, which necessitate text based communication, the scope of social network analysis only demands a relationship to exist between two actors.

For instance the importance and relevance of decision undertaken by the supreme court can be seen in this context. Any decision cites previous cases and clauses from the Constitution and the Bill of Rights. By viewing the individual court cases as actors in a network and citation of a case by the justices undertaking a decision as a link, a citation network can be constructed [46]. In such a fashion the most influential decisions and their overturning can be analyzed.

The relation used by the social network application "Six Degrees from Kevin Bacon" (namely the relationship of *X appearing in movie with Y*) is for instance perfectly usable in social network analysis. The interested reader can find this implementation of social network analysis at *"http://oracleofbacon.org/"* [98]. The distance of an actor to Kevin Bacon is termed his Bacon Number, in reference to the Erdös Number. Another instance is the example given in Figure 2.3 on page 14 which utilizes the *X is on the same board as Y* relationship.

Social Sciences Approach

The start of social network analysis can be traced to the analysis of group behavior in psychology and related fields in the 1950s. The interest was to discover individuals playing a central or leading role in small groups of individuals.

The techniques developed in this respect have been expanded to be used in contexts using any kind of relationship, not only communication.

Six Degrees of Separation

Starting in the 50s and 60s of the previous century, the social sciences focused increasingly on the understanding of group dynamics. The work done by Milgram [113] and the following work done by Freeman[51] has provided a foundation for the use of social network analysis in the academia.

The principal motivation has been the understanding of relationship and status structures in small groups, and has been developed to explore the structure of group interactions. A fundamental work has been done by [56] when exploring the type of relations existing within personal networks. This work reaches the remarkable conclusion that weak ties, such as friendships, are much more important than strong ties, such as family members or close personal friends. This work is extended by [67] into more modern aspects of cooperative work methods.

The general phenomena underlying this aspect of social network analysis is known as the **small world phenomena** [126]. In effect a social network is scrutinized to find its diameter, which even for large networks stays comparatively small [76, 122, 47]. Even the largest networks such as the World Wide Web has a diameter of just under 20 hops [5]. This is largely due to the scale-free nature of the networks, where strong communication hubs shrink the travel distance for all actors in the network.

Even more complex networks such as weblogs show such traits, which has consequences for the information transmission within these networks [58, 57]. Consequently the transmission of information is funneled through a select few actors forming a “backbone” to the community.

The discovery of such communities in larger networks has also been an area of interest. The use of network analysis to discover communities of interest [100, 45, 70] has brought insight into the cropping up of such groups on the Internet. The observation of such communities over time [104] has also brought an evolutionary aspect into network development.

Perhaps the most widely known offshoot of social network analysis is its use in indexing the World Wide Web. The HITS [77, 15] algorithm formed a first step towards automatically ranking web pages by their popularity, or linking density. The PageRank [19, 95, 59] approach then perfected this into the most widely used search engine today, Google.

Equivalent approaches have been applied to trust networks, or networks in which the relation is best described as *X recommended to Y*. Such networks [136, 138] are the foundation of the analytical software recommending products to users of Amazon.

Collaboration Networks

The application of social network analysis to works outside of the social sciences started appearing in recent years first in the context of collaboration networks. A problem faced by any scientist publishing today, it is often a long and wearisome aspect of viewing all published relevant material in a field. While mechanisms such as peer publishing and content gathering in prestigious journals makes life easier, the possibility of missing publications has increased significantly in recent years.

The development and analysis of such collaboration networks has been an attempt to establish a comprehensive view of the publishing networks, in order to facilitate the designation of important and/or peripheral publishing entities[29, 88, 28].

Un-managed Personal Networks

One category of personal networks has been the rise of what is today loosely, and arbitrarily, termed "Web 2.0". The use of the Internet as a collaborative medium encouraging participants to enter and maintain their own networks has been on the rise.

Examples of this class of social network are business contact networks such as the services offered by Xing(OpenBC)[61], or personal contact networks such as Friendster, Orkut[22] [103], or the quirky such as Dogster (Friendster for dog owners).

An environment making recent headlines is Second Life [20, 66], whose game play concept, like all Massive Multiplayer On-line Role Playing Games, has a built in reliance on creating personal contact networks on-line.

Whether the communities are in games or blogs, there has been an interest in exploiting these communities for commercial purposes. Especially the precise placement and marketing has been a target of the efforts of [99, 36].

One persisting problem these communities pose to the exploitation of network structure contained within is the heterogeneous nature of the network. In most web-based communities the cost of joining is low, leading to the misuse of relationships to such an extent that the collection of references has become a sport unto itself, thus diluting the expressiveness of social network analysis [1].

Generated Personal Networks

On the other hand an increasing trend and/or market niche has been the use of social networks analysis to provide information management and retrieval as well as management support in a business environment. The practitioners of this form are firms such as LinkedIn [63] and Spoke. An interesting proposal has been the TopicShop [8] system, intending to use content to find an appropriate actor within a social network[135].

Knowledge management has also tried to leverage social network analysis, either as a way to establish importance within a document management system[37], or to more effectively perform knowledge sharing [54]. It has also been suggested that the process of system design can be improved by designing patterns more amenable to social network analysis, or to perform better under criteria of social behavior[41].

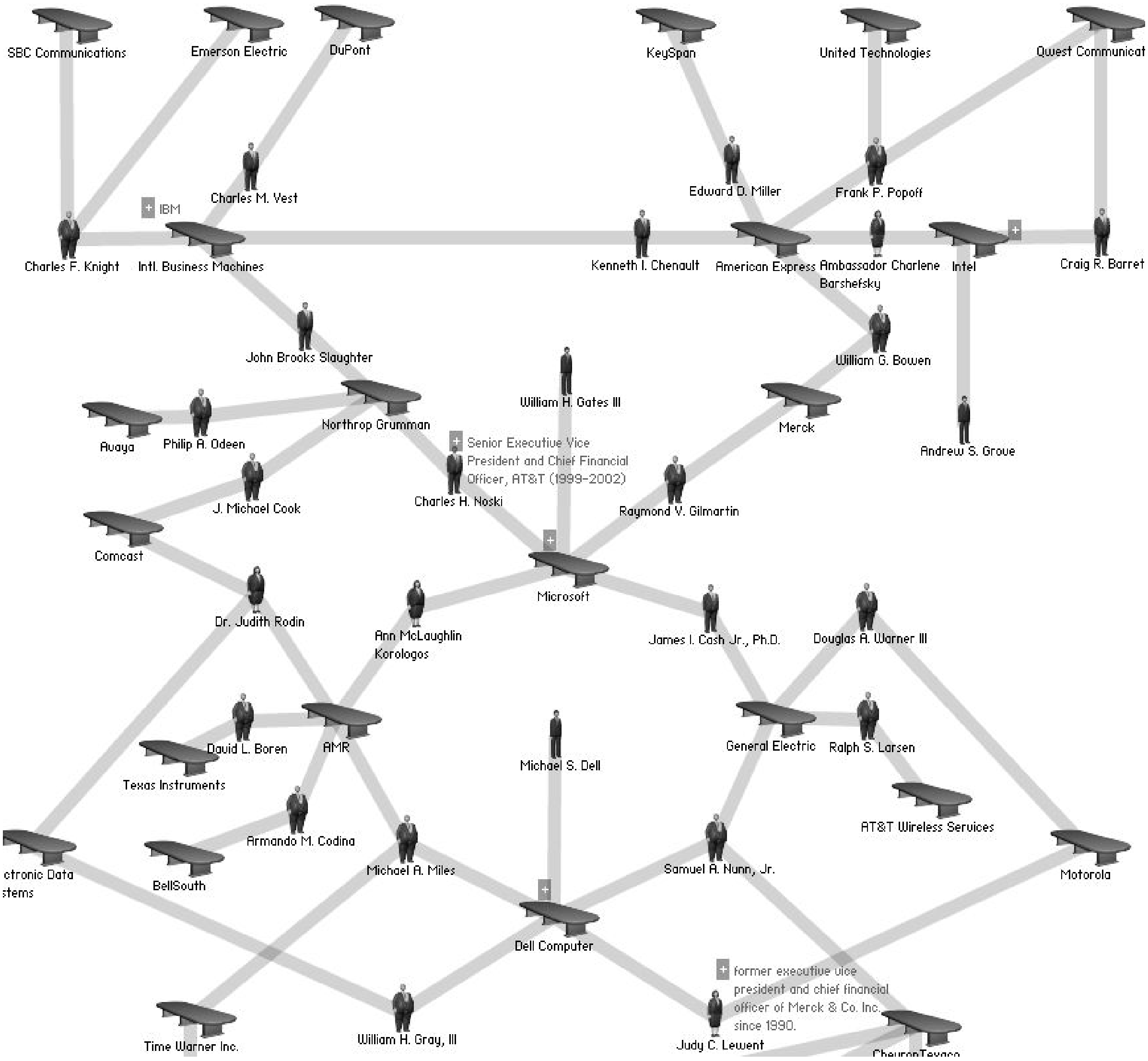


Figure 2.3: A corporate social network[94]

Relationship Intelligence

The use of social network analysis has also impacted on the use and ubiquity of social predictive modeling. The search and recommendation engines of Amazon, and the ranking of Google rely on the identification of relevant and important pages as perceived by the community at large. For instance the analysis of the interlinked nature of the Fortune 500 companies boards can lend an insight into conflicts of interest in the corporate world. This idea is illustrated in Figure 2.3. The figure shows a two-mode network, a bi-partite graph of board members associated with executive board of companies. This shows an idea of the confluence of interest across major corporations at the highest level.

2.2.2 Social Network Analysis

This section focuses on the methods and approaches used to analyze social networks today [121]. The rise of social network analysis in the field of computer science owes much to the spread of the Internet and the increasing reliance upon this medium for communication. Email alone has increased in importance from being a feature to a necessity to a critical business tool over the last 15 years.

Against this background it has become increasingly obvious that the use of email networks on the one hand, and the use of similarly linked structures in the Internet can be analyzed with the help of social network analysis.

Among the number of applicable data available one must count

- Email networks
- The World Wide Web link structure
- Peer-to-Peer networks
- Information flow networks
- Recommender networks (Referral Systems [104])
- Client maintained content networks (Wikipedia[96], User Groups[100])

These are just a few of the structured networks found today. Another impulse has been given by 9/11, after which research has been geared towards the specific problem of detecting small groups within large networks, such as detecting terrorist cells [78, 131, 92] and classification by association as well as link prediction [83].

The spread of malware of viral attacks in computer networks has also been a point of research in recent years. The field of virtual epidemiology [10, 111] takes some pointers from biology and treats the spread of viral attacks like a biological contagion. Interestingly, the spread of a computer virus often coincides with the spread of information in social networks[102].

Generally the robustness and strength of networks with respect to electronic attack [6] has been an interest in recent years. This ranges as far as the formal treatment of information warfare [123] and the use of network analysis in intrusion detection [134].

Properties

A social network has a few central properties revealing information about the structure of the network itself, and the members of the network. A few network types have been touched upon in section 2.1.2.

Properties of complex networks [108, 90] and random networks[62] have great impact upon the analysis thereof. Especially a search performed in a complex network [133] is greatly dependant on the completeness of the data at hand.

Actor	Cent.	Position
Andy Zipper	0.088335	Vice President Enron Online
Louise Kitchen	0.078280	Founder EnronOnline
John Lavorato	0.076380	CEO of Enron
Barry Tycholiz	0.071972	Vice President
Tana Jones	0.055695	N/A
Mike Grigsby	0.047822	director of corporate strategy and development.
Richard Sanders	0.047530	Vice President and Assistant. General Counsel for Enron Wholesale Services,
Geoff Storey	0.045056	N/A
David Delainey	0.042674	CEO, Enron Energy Services.
Michelle Cash	0.037957	Assistant

Table 2.1: Centrality in the Enron email network

Centrality

The centrality of an actor within a network attempts to capture the importance with respect to how much this actor is involved in the communications activity. This centrality can be defined in various ways [128, 105]. Table 2.1 shows a sample listing of the centrality values of actors in the Enron network (see section 4.1 on page 54).

- In/Out link [121]

  Basically a measure based on the number of links going in as well as out from an actor. The more links he has, the more connected and embedded within the network he is. While it works well on small networks, it is intrinsically local and takes no account of transitivity in the graph.

- Random Walk [89]

  An idea put forward to alleviate the necessity of combing through the entire graph, it proposes to measure centrality as a function of how many random walks through the graph hit a given actor.

- Betweenness [121]

Probably the most common measure of centrality implemented deems an actor to be central to the network in direct relation to how many paths it lies in between all other pairs of nodes.

- Rank [121]

  Can be taken as an extension of the in/out link centrality idea, only encompassing transitivity. Thus indirect communication to nodes further away in the network are also considered to add to the centrality of an actor.

- Eigenvalue [68, 69]

  As a graph can be represented by an adjacency matrix, and an adjacency matrix can yield real Eigenvectors. This is the case when the communication is encoded as imaginary values in the matrix, as well as being hermitian. The resulting (real) Eigenvectors can be interpreted as the main conversations, with actors ranked by importance within a conversation.

Prestige

Another important aspect has been the recognition of the level of influence a member of the group has on the whole, designated as prestige. While centrality can be seen as a measure capturing how involved an actor is in a network, prestige aims to capture his importance.

- In-Link

  Contrary to the measure of centrality mentioned above, this measure aim to capture how many actors refer to a given actor as a measure to how much he is sought after. The more in-links, the more prestige (or perhaps expertize) he has.

- Rank

  Similarly, the determination of prestige under consideration of transitivity yields the Rank prestige measure. As such it is more comprehensive and sensitive than the localized version above.

Considerations of prestige have been given increased thought in recent years, especially when considering how to increase influence through the maximization of rank in a network [75]. This basically represents the drive for importance found in much blogging and Internet based communication behavior today.

Graph Size

The intention under which social network analysis has been developed has been the use and observation of a distinct group of limited size. This has several implications which will become central to this thesis. Especially the detection of denser sub-graphs within a network has gained relevance in recent years.

Cliques

The definition and subsequent detection of subgroups within a group of people has become known as clique detection. The current methodologies focus on discovering complete or nearly complete sub-graphs within a network. The definition off a clique originates from the observation of small subgroups within a population in which every member is in contact with every other member. These complete sub-graphs form a core group of social interaction of interest to the research done.

Today the intention behind the definition of a clique can be seen somewhat more relaxed, as the completeness of a sub-graph is often more a hindrance than a help. The broadening of clique to include actors in contact with each other on a more intense level than the graph at large falls more into line with communication patterns observed today.

In chapter 3.3 on page 36 we will be introducing alternate ways to find such subgroups. Another point is the fact that the size of typical networks has increased to the point of diluting the original intention of social network analysis. It has become necessary to either change network analysis, or preprocess the data to restore social network analysis meaning and functionality.

Limitations

There is a limit to the expressiveness of social network analysis. This thesis considers the interpretation of results from social network analysis to be of great importance when conducting analysis of larger (i.e. massive) graphs. This central aspect od social network analysis wihtin the context of massive graps is explored in section 3.3.

2.2.3 An Algorithmic Perspective

So what does this mean from an algorithmic perspective? Generally all operations are performed on adjacency matrices, thus framing the dimension to be considered in the efficiency calculations.

A number of approaches have been put forward to deal with adjacency matrices efficiently and effectively, such as the detection of networks[91] in blogging data, or the treatment of social networks in conjunction with content [82].

Graph Mining

The subject of graph mining, while relevant in this context, will be treated in section 2.4 on page 25. Nonetheless it has to be mentioned that there is a great deal of research done to extract sub-graphs from massive graphs. These approaches all bear on the subsequent analysis using social network analysis.

SNA by Social Scientists

The approaches used by social scientists are based on methodologies developed as far back as the mid- to late seventies, and as such are poorly adapted to the computational power available today.

- Block Models[127]

  are used in the analysis of social networks when the individual actors are combined into discrete subgroups. The subgroups are then linked with one or more link types, each expressing a different relationship. This approach aims to provide a slightly more generalized form of social network analysis based on groups rather than individual actors.

- Positional Analysis[21, 86]

  can be said to relate to the previously mentioned Block Models. In it the position of an actor within a network as defined by the in- and/or out-degree describes a role within the network.

- Relational Equivalence [121]

  Aims to group actors by their role within a network. Actors with equivalent structural topology are assumed to posses similar role within a social network.

For the course of this thesis we will focus on the aspects of centrality and prestige in social networks. The extension to other aspects of social network analysis is an area to be explored by future work.

Centrality/Prestige algorithms

Chapter 5 will mention the implementation of the basic algorithms. While all options mentioned in this section have been implemented, we concentrate our analysis on the use of betweenness centrality.

Eigenvector Approaches

An interesting approach to describing centrality is the observance of equivalence between the intention of centrality in a network, and the calculation of Eigenvectors of a matrix. Let the social Network $\mathcal{SN} = (\mathcal{A}, \mathcal{L})$ be described by the adjacency matrix M. An actor $a \in \mathcal{A}$ can then be said to have a centrality value based on the centrality of everyone in his neighborhood $N = neighborhood(1, a)$:

$$centrality(a) = \frac{1}{\lambda} \sum_{a_n \in N} centrality(a_n)$$

where λ is a proportionality factor. Now formulating this in terms of the adjacency matrix gives us

$$centrality(a_i) = \frac{1}{\lambda} \sum_{j=1}^{size(\mathcal{SN})} M_{i,j} centrality(a_j)$$

which can be written in vector notation as

$$\overrightarrow{a} = \frac{1}{\lambda} M \overrightarrow{a} \leftrightarrow M \overrightarrow{a} = \lambda \overrightarrow{a}$$

which is the Eigenvalue equation. Hence the centrality of a node is proportional to the Eigenvalue distribution of the Eigenvector solution.

This approach can be extended to use a bidirectional coding of communication. When encoding the number of messages sent and received as an imaginary number and rotating this hermitian matrix to attain only real number, the Eigenvector solution is not changed, but expresses centrality of the complex adjacency matrix [68].

The drawback of this elegant method is the computational complexity used to calculate the matrices and the Eigenvector solutions. While analyzing a group of several dozen is feasible, the analysis of several hundred or even thousands of actors is quite out of the question (see chapter 5).

2.3 Text Mining

The second field this thesis makes use of is the domain of text analysis [84], or more specifically the use of topic discovery in text based messaging corpora. After the introduction of network analysis in the preceding sections, the aspects of topic discovery pertinent to this thesis will now be outlined.

2.3.1 Semantic Content

A current and evolving field of research is concerned with the handling and processing of unstructured text. This approach has been chosen for this thesis as the data gathered is mostly unstructured to a great degree. Email or blogging networks rely on messages being passed back and forth without any kind of semantic markup, mostly even bereft of the touch of a spell checker.

The first challenge faced by computational processing of unstructured text is the recognition of content. In order to overcome the inability to recognize and understand content, the foundation of text mining is the assumption that structure follows content. Particularly in respect to topic discovery, i.e. grouping documents with similar content, the assumption that similar word occurrences indicate similar discussed content lies at the heart of the process.

While the field of text mining is large, we aim to follow one particular goal; namely the discovery of topics within large text based message corpora.

Capturing semantics

While the field of text description is vast in itself, there are a few aspects of particular relevance to this thesis. These include:

- Content extraction

  The basis for any analysis using a keyword based approach needs at first to extract said keywords. The process itself is straightforward, indeed a simple parsing exercise, save for a first differentiation in importance. Common words conveying little concrete meaning, such as the word *'the'* or prepositions and adjectives in general, can be argued to contribute little to the content of a text. The process of content extraction followed by this thesis utilizes three aspects:

  – Porter Stemming
  – Stop Word Lists (Blacklists)
  – Dictionary based spell checking (White-lists)

- Keyword Vectors The occurrence of keywords in a message can now be taken to form a vector precisely describing the content of a message. The content space is constructed over the number of stems occurring in the document corpus. This representation has the desirable quality of being easily processed in an automatic fashion.

- Discarding too frequent/infrequent stems It is often beneficial to restrict further processing to such stems which do not occur to often or too seldom in the data set. Too frequent an occurrence robs the stem of meaning while too infrequent of an occurrence does not add enough information to warrant an entire new dimension in the keyword space just to satisfy an insignificant handful of messages.

  One large problem encountered during the course of this thesis is the observation that the rough shod discarding of too many infrequent terms can adversely impact the further analysis. The cause of this lies with the distribution of words used in text, where an overwhelming majority of words has a very small occurrence in written text [38].

- TF/IDF [4] The keyword vectors counting absolute occurrences of stems have a significant drawback. They describe a discrete space populated very sparsely by individual vectors. Section 4.1 and section 4.2 wind up having huge vector spaces with relatively few vectors occupying them. For this reason the vectors can be normalized to reflect the relative importance of individual stems. By weighting the occurrence of a stem by the frequency it appears in the entire corpus with respect to the inverse of the frequency it is found in a document, important terms describing content are bolstered while stems contributing little are devalued.

  The weight of a term is inferred by a term frequency within a document in relation to the inverse frequency of documents containing the term within the corpus. Thus the shorthand of term frequency in relation to the inverse document frequency. The

Data Set	Dimensionality	Size of Data Set	Avg. Occurrence of stems per Dim
Enron	7856	491754	44.50535
Blogging	12502	3950	5.25745

Table 2.2: The dimensions used to describe data corpora in section 4

use of this method has allowed the significant reduction in vector space through the use of singular value decomposition in section 5.3.2 on page 92.

By using words, which are readily parsed by machines, the capturing of information becomes possible on a large scale. There are formidable problems with this approach, such as the variation of spelling, and the recognition of proper names, which are particularly significant in unstructured texts such as email.

Another restriction has been the loss of integrity of a given communications network based on the fact that many small email messages are mainly generic and are filtered out of the data set. The more such communications are dropped from the analysis, the sparser the adjacency matrix becomes.

The extraction and use of keyword vectors describing content is an ongoing field of research. In practice the use of identifying keywords is an arduous and complicated task, often more an art than a science.

High dimensional space

As mentioned previously, the keyword space describing the contents of a document corpus occupies a increasingly large number of dimensions. The corpora under consideration in the latter section of this thesis are described in table 2.2.

A point of this exercise in capturing text content is the determination of similarity between messages. By use of keyword vectors it becomes possible to compare text numerically by using distance measures operating within this keyword space.

But the large number of dimensions involved in keyword spaces capturing content in text corpora presents difficulties in further text analysis. It is increasingly difficult to gain insights into message clusters on account of the curse of dimensionality[44]. This fact is exacerbated by the sparsity of data within this space; it becomes increasingly difficult to find any kind of content groups within this data.

The unsupervised analysis of text based messages lies at the heart of this thesis. It has therefore been of great concern to explore the possibilities offered by text analysis approaches in processing content and interpreting similarity between messages. This thesis has explored two basic approaches to the problem, both being delineated in the following subsections.

Text Analysis

The development of information retrieval techniques for use in the field of text analysis [84] is a well explored and ongoing new area of research. This thesis has utilized many features of text analysis to extract semantic content from unstructured text.

The ultimate aim of this thesis is the understanding of a text based message corpus under the specific aspect of topics. In order to extract topics from such a corpus, the similarity of messages based on similar content must be performed. Two basic approaches to the understanding and utilization of content have been looked at.

Probabilistic Analysis

By interpreting keyword vectors not in a spatial context, but as a probabilistic expression of individual messages, the correlation between messages can be seen in the light of probabilistic closeness [64, 34]. By comparing relative occurrences across messages and weighting individual terms according to statistical relevance in relation to the entire corpus [38] it is possible to establish semantic similarity.

A common measure used in this case is the Kullback-Leibler Divergence [79], while not being a metric, provides an asymmetric measure generally interpreted as distance between two probabilistic keyword distributions.

Clustering

By interpreting keyword occurrence in messages as spatial vectors the application of spatial clustering methods come into the purview of analysis techniques. As the following sections will remark, the use of vector representations of messages lends itself to the spatial interpretation of topics as densely packed vectors in space. While spatial clustering based on nearest neighbor techniques [52] have been around for some time, this thesis will concentrate on density based clustering techniques [43, 9] as the most flexible and powerful approaches available.

To this extent the use of density based clustering has found application during the course of this thesis. As mentioned previously, the extraction of keyword vectors to describe content generates a high dimensional vector space populated by the individual texts, or in this case messages. The level of semantic overlap can now be quantified by using an appropriate distance measure.

There is a veritable zoo of density based clustering algorithms, of which a select few have been implemented in the course of this thesis. The density based clustering algorithms used are DBSCAN [42], OPTICS[9]. Not to leave out the class of hierarchical clustering algorithms, a version of the Single Link[101] approach has also been implemented (see chapter 5). This approach has been followed since this very basic clustering algorithm is simple to implement and reacts predictably to high dimensional data. As the results were unsatisfactory in comparison to the density based approaches just mentioned, a third class of algorithms have also been used. Algorithms from the partitioning side have been included for comparison purposes, the perennial classic K-Means [85] and slight derivatives [71] in this case.

With the help of clustering techniques it is possible to extract groups of messages with generally similar content, thus discovering topics within the data corpus. The performance of these algorithms unfortunately decreases when data are sparsely distributed in a high-dimensional space. For a discussion of the computational aspects when reducing the number of dimensions the section 5.3.2 on page 92 goes into greater detail.

2.3.2 Content Structure

By being able to analyze and group messages into semantic clusters, it is now possible to equate clusters of similar content with topics. The use of density based clusters allows a good grasp on the semantics of messages, and the classification of noise aids in the creation of clusters without the background noise of irrelevant messages.

Topics

Topic detection has been a field of research enjoying a fair amount of attention in recent years [7]. The stated goal of finding clusters of documents with similar content will be used as the basis for the first of two approaches to segment networking graphs for social network analysis. Topics will be used in chapter 3.3 on page 36 to segment the data corpus into semantically coherent sections.

The process of topic detection is computationally intensive when confronted with massive graphs, and as such has motivated the exploration of topics in a more general framework. For instance, the complexity of the process used to detect topics becomes increasingly unwieldy when looking at messaging corpora such as the Enron data set (see chapter 4.1).

Topic Hierarchy

Topic clusters are not naturally completely homogeneous. By applying techniques to recursively extract ever smaller clusters from a topic, it is possible to construct topic hierarchies. The discovery of such hierarchies can be of great aid when the initial clustering is performed in a very loose fashion, in order to find extended largely inconclusive clusters (such that there might not even exist a single topic per se, but more of a theme).

These topic hierarchies, from general to specific, can be used analogously to the topic clusters discovered in the initial, flat, approach. This methodology has not been followed in this thesis, for there exists another approach to creating topic hierarchies, based not only on a semantic hierarchy, but including a temporal aspect into the construction of the topic tree [116].

Topic Tracking

Text corpora are comprised of messages spanning diverse themes, which themselves can be seen as broad general topics or more specific but tightly defined local topics. For instance the general theme of reporting on technological advances contains various strands concerned with the development and launch of the Apple iPhone. The exploration of such content structure and topic hierarchies has been well explored in recent years. The goal of this work is to focus the use of topic discovery on the observation and tracking of the evolution of topic hierarchies. We want to not only find the topic structure of news on the iPhone, but its place within the trend of mobile phones in general, and the development and transmutation over time of a given topic.

The need to analyze and track news topics over time has been an area of interest to Siemens for some time. The approach described herein is entering the prototype phase as we speak and will be rolled out in the near future. We expect to extend corporate services by developing a support tool offering methods to locate and apprise new topics and particularly trends over an extended period of time.

While the information contained in individual messages and announcements changes rarely to never, the text corpora comprised of many such messages are never static but rather evolve over time. Modeling temporal evolution through the use of sliding windows is a common approach of temporal analysis of text corpora. Whereas it gives adequate insight into adjacent or overlapping time periods, it lacks an understanding overarching several such time periods.

We therefore focus on observing the development of topics over time; it is of interest to discover which topics gain attention and which disappear into oblivion. Modeling these developments over time lends us the tools to track and analyze topics in a method independent of time slices by themselves, and perform analysis more attuned to the individual nature of a topic in consideration with its life-cycle.

Topics detected in news corpora, such as the coverage of the iPhone or a market correction on Wall Street, are inherently tightly localized in time. A specific topic might only be discussed for a few days or a week (the iPhone must be launched sometime) after which this topic will not exist in the media anymore. But any discussion of a tightly defined scope is embedded in a more general trend, and other topics will be related to this one simply by belonging to this same trend. A discussion about the impact of the iPhone on Apple is surely linked to the coverage of its release. We want to find and track the evolution of these trends, and the approach is described in detail in section 4.3.

2.4 Graph Mining

Where the previous section has focused on the analysis of the content provided by a text messaging corpus, there is of course a second aspect to be scrutinized: Every communications corpus can be analyzed on a structural basis. The dissection of graphs has been implicitly used by social network analysis since its inception.

The subject of graph mining is a recent addition to the data mining field. Having found entrance to data mining by the proliferation of graphs and graph based structures found in all things related to the development of the Internet and the World Wide Web, graph mining has become a valuable tool in network analysis.

The use and relevance to social network analysis is twofold. Firstly the detection of cliques in graphs is a problem long considered and not addressed satisfactorily in the classic social network analysis. The second aspect is the use of graph mining techniques to supplement and diversify social network analysis.

2.4.1 Cliques and Cohesive Subgroups

It has been a primary aim among the uses of social network analysis to extract interesting information from structural aspects of the communications network. This leads to the use of various techniques to find and extract subgroups within the entire network.

Social Network Analysis possesses a number of tools to define subgroups. These are derived from the original intention motivating the analysis itself, namely the observing and understanding of groups of people. As such the notion of cliques and similar "strong" social subgroups underlie their definition.

It is the intention of this thesis to utilize the finding of subgroups to present a second way of segmenting massive communication corpora. As can be seen in chapter 4 on page 53, the use of structural analysis preceding to semantic analysis can provide a crucial reduction in processing time.

Cliques

Social network analysis has developed the notion of a clique from the observation of small subgroups of observed populations to exhibit a more intense communication than the group at a whole. The discovery of such groups and subsequent study has then lead to the adoption of this most restrictive definition of a clique.

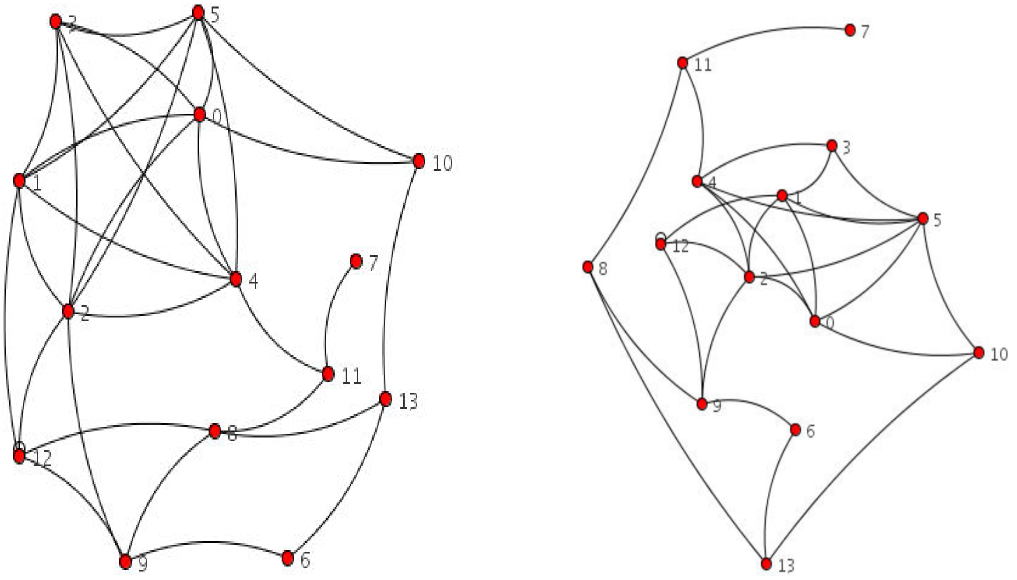
Cliques have various forms, depending upon context. Due to the varying use of the term no general definition can be given, but a concept of use can be defined as follows:

Clique

Let $\mathcal{SN} = (\mathcal{A}, \mathcal{L})$ be a social network with a small subgroup $\mathcal{C} \subset \mathcal{A}$. Iff this subgroup $\mathcal{C}$ is more strongly connected than the entire graph, it is deemed to form a clique within the entire group.

The precise understanding of *more strongly connected* depends on the various forms of cliques considered. The most direct way is to demand a maximal sub-graph, i.e. each member of a clique communicating directly with every other member of the group. This definition evolved in an environment where the groups observed rarely exceeded a few dozen actors, and therefore warranted such a restrictive approach. It is apparent that this notion, while being direct, is of limited applicability in the context of real world data and massive graphs.

Therefore we will consider two of the more generic forms in use. These terms are cast in a more lenient form than the previous clique definition, and are collectively referred



(a) n-Cliques

(b) k-Plexes

Figure 2.4: Clique Variants

to as *cohesive subgroups*.

n-Cliques

Instead of demanding completeness of a sub-graph, the condition of being strongly connected can be relaxed somewhat to require a lower bound on the geodesics within the sub-graph. By assuring a minimal shortest distance between any two actors in the sub-graph the idea of being strongly connected is preserved, while at the same time introducing some leniency in relation to the upper limit of communications between actors.

n-Clique

Let $\mathcal{SN} = (\mathcal{A}, \mathcal{L})$ be a social network. $\mathcal{C} \subset \mathcal{A}$ is considered an n-Clique iff $\forall a, b \in \mathcal{C}, a \neq b : d(a, b) \leq n$.

This goes some way in addressing the inherent limited scope of cliques, but is limited by the fact that a geodesic is not constrained to the subgroup itself. To overcome this weakness a number of extensions to this approach have been proposed. Since we do not use these as such, a comprehensive discussion can be found in [121]. The use of n-clubs is introduced, where the geodesics considered are not limited to the sub-graph itself.

Figure 2.4(a) shows an example of a n-clique. There exists a clique of six people, extended to a group of 10 as a 2-clique.

k-Plexes

Another approach is to define a subgroup not solely by the link structure, but to use density as a motivation for forming sub-graphs. This definition is more closely aligned with the intention expressed in this thesis, namely the discovery of sub-graphs exhibiting a higher communications density than the rest of the graph at large.

k-Plexes

Let $\mathcal{SN} = (\mathcal{A}, \mathcal{L})$ be a social network. $\mathcal{C} \subset \mathcal{A}$ is considered a k-Plex iff

- $|\mathcal{C}| \geq 1$
- $\forall c \in \mathcal{C} : neighborhood(1, c) \geq k$
- $\forall c \in \mathcal{C} : \exists d \in \mathcal{C} \land \exists E(c, d) \in \mathcal{L} \land neighborhood(1, d) \geq k$

Figure 2.4(b) shows a graph consisting of a clique containing six members being extended by 3 members when a 2-plex is considered. Of course every member of clique is also a member of the 2-plex. The extension of a clique by adding member with a sufficiently high communication density has shown itself to be more stable and flexible than the clique approach [121].

In contrast to finding such cohesive subgroups in communications graphs, the use of data mining approaches to graphs has a slightly different focus.

2.4.2 Sub-graph Extraction

The introduction of data mining to graphs has the intention of finding previously unknown yet interesting information in data described by graph structures [24, 120]. The use of graph mining takes various forms, a comprehensive overview can be found in [32].

Graph mining comes in various shapes and sizes, but can be broken into several areas:

- Finding frequent sub-graphs
  has been the goal of approaches trying to find similarity within graphs. The goal of this approach has been the discovery of isomorphic sub-graphs. This approach has been exemplified in the SUBDUE [65, 31, 72] system.

- Finding dense sub-graphs
  Of far greater interest to this thesis is the discovery of sub-graphs exhibiting a higher communication density than the surround graph it is embedded in. This idea is followed by a number of approaches:
  - Geodesic Clustering [130] and Graph Clustering [35]
    By observing that a structural property of denser sub-graphs is mirrored in the increased existence of geodesics running through its nodes, this approach searches for sub-graphs using a heuristic observing and following geodesics. Graph clustering is based on the notion of finding and building areas of nodes with a high concentration of links.

- Shingling: Large sub-graphs in the World Wide Web [53].
  In contrast to the previous approach, this approach uses content to find subgraphs of similar content. The term shingling expresses the use of keywords to express similarity between nodes. Should a node have neighbors who have a minimal overlap of keywords with the node under scrutiny, then the subgraph is expanded to include these adjacent, "similar", nodes. The definition of overlap is by nature flexible, as the similarity can drift from one node to the next.
- Dense Sub-graph Discovery
  postulates the feasibility of clustering massive graphs using the existing approach of density based clustering methods. Using the distance between nodes as a measure of comparison, any graph can (quickly) be clustered. There is a prerequisite, namely the need to be able to work in a non-euclidean space. A graph has absolute distance between nodes (which can reach infinity), but has no intrinsic spatial structure. But since any two nodes within a graph have a distance specified by the shortest path, a densety based clustering approach can be said to exist. But any clustering algorithm needing a spatial representation with vectors is unsuitable for this purpose. For example, K-Means relies on the calculation of a euclidean mean vector not available in the context of a graph. On the other hand, the class of density based clustering algorithms relying only on a distance measure can function very well, as the euclidean distance measure is replaced with a graph-kernel based distance measure.

2.4.3 Density Based Graph Mining

The idea of general applicability of density based clustering techniques to dense subgraph extraction has found application in this thesis. As can be seen in chapter 5 implementations of such algorithms have been used to extract sub-graphs from massive graphs.

The dense sub-graph extraction method used in this thesis is the replacement of the euclidean distance measure to measure distance in a graph. The distance measure can be applied to graphs when the following conditions are met:

- The target clustering algorithm does not rely on a euclidean distance measure, and does not need any direct information from a euclidean vector space during the clustering process.
- The graph is sufficiently large.

We have implemented and tested this approach by clustering the communications graph of the Enron data set using the OPTICS algorithm. The example can be seen in the case study in chapter 4 on page 53.

3 Partitioning Massive Graphs

Contents

3.1 Introduction

In a world increasingly reliant on electronic communication, businesses and corporations today have a powerful incentive to attain a better understanding of what occurs within modern communications media. Be it either to monitor the pulse of their own business, or to keep track of the perceived image of themselves by consumers or just to keep abreast with evolving trends, a bountiful source of data to base such decisions on is only an arms length away. The Internet and all technologies based on communication contain opinions, valuations, and procedural knowledge on almost every larger business today.

Finding such information in the vast amount of communications data has become a challenge in itself. The sheer size and magnitude of electronic messaging poses its own problems to the effective and timely analytical grasp of day to day business decisions.

This issue lies at the heart of the thesis as well as this chapter, and we focus on several methodologies geared toward the extraction of relevant data from large communications corpora. By utilizing two aspects of the data available, namely the semantic as well as the structural side, we can bring a number of approaches to bear:

- **Network Segmentation** uses the content of a communications network to slice the network into multiple layers according to distinct content. This approach generates a number of semantically cohesive sub-graphs from a heterogeneous all-encompassing communications graph.

- **Content Based Social Network Analysis** uses semantically cohesive sub-graphs to imbue the social network analysis with greater meaning by focusing and constricting sub-graphs to common content. Not only can social network analysis be conducted within such semantically homogeneous subgroups, but between them as well.

- **Density Based Network Segmentation** starts from a structural perspective by grouping neighborhoods of high density within a graph together. A neighborhood of denser link placements yields structurally focused heterogeneous sub-graphs.

- **Segment Evolution** tracks the evolution of semantically cohesive sub-graphs over time and adds a temporal aspect to content based social network analysis.

Toady's communications media are transporting ever more information. We motivate the development of more discriminating social network analysis in the following section with the loss of precise meaning when analysis is performed on increasingly large and diverse communications networks. Each approach will shine a light on a different aspect of mining massive communications corpora. Chapter 4 will show the increase in meaning when social network analysis is performed with or without semantic targeting.

3.2 Massive Corpora

After covering the fields of text analysis, graph mining and social network analysis as pertains to this thesis in sections 2.2 through 2.4, there remains a discussion of the data this thesis aims to explore. As the title suggests, the analysis of massive communications corpora are the target of this thesis.

Many kinds of communication networks, and we list a number of relevant types on the next few pages, have been increasing in scope and activity in recent years. The more people partake in email exchanges, and the larger the communities grow, the more traffic is generated. These massive communications corpora can be captured by communications graphs, which are not susceptible to the generalized application of social network analysis. This motivates the specific focus on massive graphs.

There are several sources of communication corpora considered in this work, partially to underscore the wide applicability, but also to gain a wide test-bed for the case study in Chapter 4. We believe the combination of content with social network analysis will greatly improve our understanding of real-world communication.

3.2.1 Communication Networks

The application of social network analysis to graphs found in the World Wide Web and the Internet has received increasing attention in recent years. Networks as diverse as those generated by e-mail communication, instant messaging, link structure in the Internet as well as citation and collaboration networks have all been treated with this method. So far these analyses solely utilize graph structure. There is, however, another source of information available in messaging corpora, namely content.

When communication patterns were being analyzed just a few years ago, networks containing no more than a few dozen people were the norm [48, 49]. These, by today's standards, mostly small case studies provide the groundwork for modern methods, but focused on networks gathered and prepared by hand.

In contrast to such an academic setting, automated analysis of communication patterns today face an entirely different class of data. Not only is the number of actors exponentially larger, but also the use of electronic communication has become ubiquitous. This translates into an increase in exchanged data unforeseen only a decade ago.

Massive Graphs

Today's communications networks are exploding in complexity and size. When analyzing the conversation in a room it can be enlightening to discover captivating conversationalists, but when listening to an entire town hall or football stadium, the observation of central actors becomes more and more general to the point of losing all expressiveness.

When looking at data sources coming into the purview of social network analysis today, the need to digest and analyze larger and larger networks is at the forefront of modern automated analyses. The data sets used in academia have themselves grown in size (the Enron data set used in this thesis has half a million emails), but are in turn themselves dwarfed by the real world data envisaged by such sources as email data in a large corporation. The email generated during the course of a year can easily exceed the terabyte limit, and as such is not amenable to supervised analysis.

As this thesis focuses on such data sources, we consider data which precludes interactive or manual analysis to be considered "massive". Hence the focus on such communication corpora captured by **massive communication graphs**.

While massive graphs generally are network constructs of size and extent comparable to the number of nodes in the Internet, we do not consider the size requirement to be the only component which make massive graphs challenging for social network analysis. We want to note another property of massive graphs, namely the super-positioning and inclusion of multiple, indeed many networks into a single super-network.

For instance a newsgroup can contain many threads upon a single subject, or even several discussions about similar yet subtly different subject matters. The juxtaposition of such similar but distinct networks would then yield a (albeit small) massive graph.

The occurrence of such co-mingled and massive graphs is almost ubiquitous. If no care is taken to single out different conversations, any newsgroup or even web-log contains multiple strata of discussion. Performing social network analysis upon such multifaceted data will only provide insight into general communication behavior. It is these semantically complex networks we will focus upon.

Interpreting Social Networks

It is important to note the stark difference of real world data in contrast to the requirements of social network analysis. The sheer size and extent of networks gathered in e-mail and other media is completely out of scope of network analysis.

The roots of social network analysis lies in the surveying and structural analysis of small but strongly interacting groups of actors. This has a significant impact on the relevance of measures produced by social network analysis. Centrality measures carry structural meaning only, hence it is necessary to extract centrality values from semantically cohesive graphs or sub-graphs.

Large graphs, or in the context of this thesis massive graphs, are perfectly amenable to *analysis* using SNA. But should the data contain sub-graphs of significant individuality, then the expressiveness of social network analysis decreases significantly.

Networks under scrutiny today are increasing in size and complexity, being often an amalgamation and super-positioning of multiple networks. But as networks get larger, the more general and imprecise are the conclusions drawn from social network analysis. Google will score a very high centrality, based solely on the fact it links to almost everything.

Motivation I

Therefore the intention of the analysis methods does not fall into line with the expected data found in real world problems.

This suggests the preparation of data to be analyzed into semantically coherent communication networks. There are several ways of doing this, either in the data gathering stage or as a preprocessing step undertaken prior to the actual social network analysis.

It is generally accepted to prepare data in accordance with this intention before applying social network analysis methods. The data is either gathered in a semantically defined locale, such as a single news group or forum, or is limited to a group of actors known to be in close proximity, such as people in a department or project group. These approaches are implicit and rely on a-priori knowledge, which we will want to abstract from.

We believe the extension of network analysis to include content is a conclusive and relevant next step to further understanding communication networks. The idea of network analysis in conjunction with semantic information can yield a more targeted and differentiated view of network data, precisely when large text based communication corpora are concerned.

3.2.2 E-Mail

A bountiful and in recent years ever increasing subject of social network analysis has been e-mail in various forms. In contrast to historical data and collected data sets this source of network data has been bountiful in theory, and in some cases also in practice (the Enron data set [30]). But there are numerous problems in the real world gathering and collating of information presented by this class of networks.

First and foremost among the concerns relating to e-mail is the privacy of individuals partaking in the network [132], limiting the potential for gathering data. The application of privacy laws differs starkly between different countries, for instance any emails written on a corporate computer is considered fair game in the United States, whereas the determination of what is private and should be restricted is far more difficult in most European countries.

Another problem when analyzing email networks is the inherent inaccuracies when gathering data, necessitating an effort in data preparation and cleaning [109]. The emails are written increasingly in an informal fashion, often neglecting correct spelling and punctuation as the effort is seen to be unnecessary on most short email messages. Only when the email takes on aspects of more formal communication does the content become more stringently formatted as would be expected in a handwritten letter. This has of course an adverse impact on unsupervised text mining of the content, often leading to excessive pruning of content.

Extracting networks [33] has been the focus of research in recent years. This starts with the simple syntactical parsing of the email at hand, and can go into the area of constructing links from inferred information. Such link generation can be based on both the structure of an extracted network as well as the content[83].

The analysis of email data is of course one of the prime goals of this thesis. Email data constitutes the most prolific and easily attainable data on communications networks in a corporate environment. Nonetheless other forms of electronic communication have also been scrutinized during the course of this thesis.

3.2.3 Web-logs

A second data source considered in chapter 4 is the area of on-line web logs. Blogging has achieved an increasingly central status over the past few years, having a similar intention in communication as newsgroups, and before that Bulletin Board Systems.

Unlike directed email messages, posting in a web-log has more in common with verbal conversation in which the target is informally identified by visual and verbal cues. In a weblog environment the intended target has to be inferred in order to present a complete link in terms of a graph (see chapter 2). While the posting of an initial message does not convey specific communication intent, as the post is directed at the readers at large and not at any specific actor, any replay can be taken to be directed at the initial poster[87]. Thus a web-log can be seen as a message based communication network.

The discovery and extraction of social networks in newsgroups and web-logs [3] has been under scrutiny in recent years. To this end the automatic mining [16] of web-logs and newsgroups and the subsequent discovery of structure[26], as well as the evolution of topics [80] has been under investigation.

Similar networks consisting of communications exchanges expressing sentiment can deliver a better understanding of the mood of a community [137]. This has lead to the development of brand monitoring via the extraction of social networks and observation of these networks.

3.2.4 Web Mining

Another aspect in the focus of social network analysis has been the link structure of the World Wide Web. Links between pages together with the designation of web sites as actors allows a comprehensive analysis of the links with respect to centrality and impact of web sites. On the one hand the use of social network analysis has spawned numerous applications ranging from the indexing of the World Wide Web (such as by *Google* or *Yahoo!*) using the centrality to better target responses to queries. On the other hand link structure can be encouraged to reflect centrality, as by the use of *Digg.com* and similar social bookmarking sites.

While the use of web mining to extract meaning and efficiency from web structures [106, 107, 117, 118] can be seen to operate upon such link structures, these methods can be extended to use social network analysis to align the structure of websites to better fit the users expectation. This avenue of research is beyond the current scope of the thesis, and will be explored as future work.

3.3 Interpretation of Social Network Analysis

When faced with massive communication graphs as described in section 3.2, a basic decision regarding the exploration of said data must be taken. Effective yet time consuming manual preparation of data must be weighed against the more efficient yet often less effective unsupervised mining of the data corpus. While tools exist to tackle different aspects such as content or structure (covered in sections 2.3 and 2.4), the combination of these holds the key for a better analysis of massive text based messaging corpora.

We propose a way to segment the data into useful slices. When focusing on large text based communication corpora, we are able to perform semantic analyses based on content of the messages, and tie the content to individual topic based sub-networks of the overall communications network. This **topic based segmentation** of a communications network yields a more differentiated understanding of relationships within the network than a more general analysis of the entire network.

While social network analysis can extract the relevance and impact of actors within a network from structural considerations, it does not regard content of the network under scrutiny. Thus the consistency of the network has a great impact upon the interpretability of the results of social network analysis. The semantic context of a network can imbue social network analysis with an added value through greater relevance in view of this content.

The expressiveness of centrality and prestige scores is determined upon the exact definition of the network analyzed. Motivation II

Comprehensive and large networks enable a judgment on the overall picture, but not a finer weighing of actors within that network. As only the structure of communication networks is considered, this structure must follow the intention of the analysis in order for social network analysis to be expressive.

The point of social network analysis is the discovery of relationships and relevance of actors in a semantically cohesive network. This assumption of viewing a single semantically cohesive structure is not always the case. Massive graphs provide ample opportunity for deviations and discovery of multiple communities.

It is important to remember that social network analysis is effectively a structural analysis. It gleans information from a purely graph based point of view. The network construction, in other words, determines the expressiveness and interpretability of the results obtained from social network analysis.

The observance and analysis of large and complex data sources using social network analysis therefore is limited to a very stark degree to general observations of network structure. Any analysis must perform a preliminary step of filtering the network with the aim to provide greater meaning to the results of social network analysis.

3.3.1 Treating Massive Graphs

Analyzing large communications networks typically includes an implied filtering or classification step in order to focus on meaningful sub-networks. The principal techniques can be summarized as implicitly limiting the data in the cleaning stages of the analysis.

While massive graphs generally are network constructs of size and extent comparable to the number of nodes in the Internet, we do not consider the size requirement to be the only component which make massive graphs challenging for social network analysis. We want to note another property of massive graphs, namely the super-positioning and

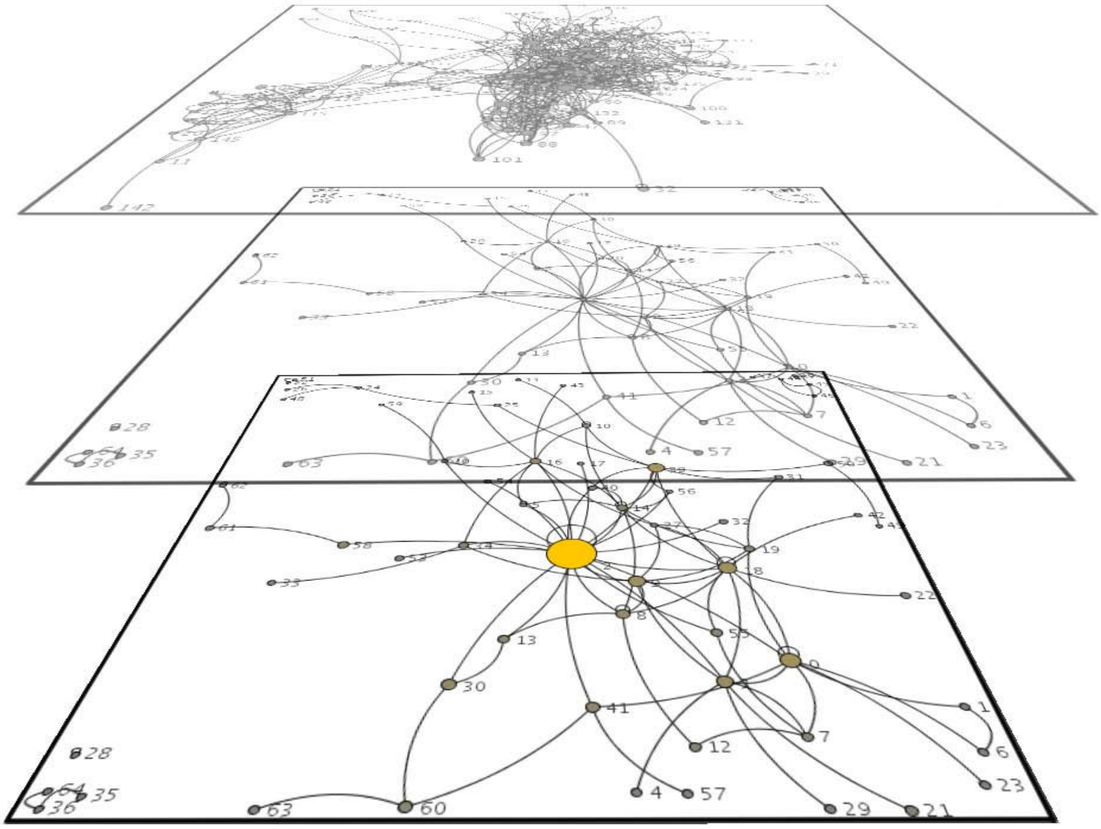


Figure 3.1: Spectral Analysis

inclusion of multiple, indeed many networks into a single super-network.

As described in section 3.2, massive graphs consisting of a number of sub-graphs form the basis of our considerations. When looking at massive communication corpora we can expect any forum to contain many threads upon a single subject, or even several discussions about similar yet subtly different subject matters.

The occurrence of such interleaved sub-graphs is almost ubiquitous. If no care is taken to single out different conversations, any newsgroup or even web-log contains multiple strata of discussion. Performing social network analysis upon such multifaceted data will only provide insight into general communication behavior.

In order to receive expressive judgments from social network analysis it becomes necessary to treat the network under scrutiny prior to analysis. This has been implicitly done in the data preparation stage, whenever the data to be collected is determined to fit into a definition of relevant communication content. This definition of relevance forms a constraint upon the data to be analyzed.

Apriori Constraints

To achieve the aforementioned cohesiveness for analysis, one can use apriori knowledge to determine how to fashion the communication network as a preprocessing step. By restricting the view of data gathering to the scope of a single news bulletin board, or the selection of a single messaging channel the assumption of selecting messages concerned mostly with a described topic is injected into the data.

It is clear that the use of such impositions of constraints upon the data is useful, but it does not facilitate the envisioned goal of unsupervised mining techniques.

Structural Constraints

Any communication network, seen as a graph, can be analyzed with the help of clustering techniques designed to find denser or more connected sub-graphs within the whole graph. The denser sub-graphs, or cliques, yield readily to social network analysis. Any such connected group of actors falls within the prerequisite context used to determine relative centrality within a group.

While this is clearly the case, we want to point out that this approach completely disregards **what** these actors are actually talking about. And while the analysis of such a clique will readily yield which member is the leader of the pack, it may rest upon the fact that one member is respectively more verbose than his or her colleagues, while a very concise but intermittent communicating actor will fail to register with a comparatively sparse ego-network.

3.3.2 Mining Graphs for Constraints

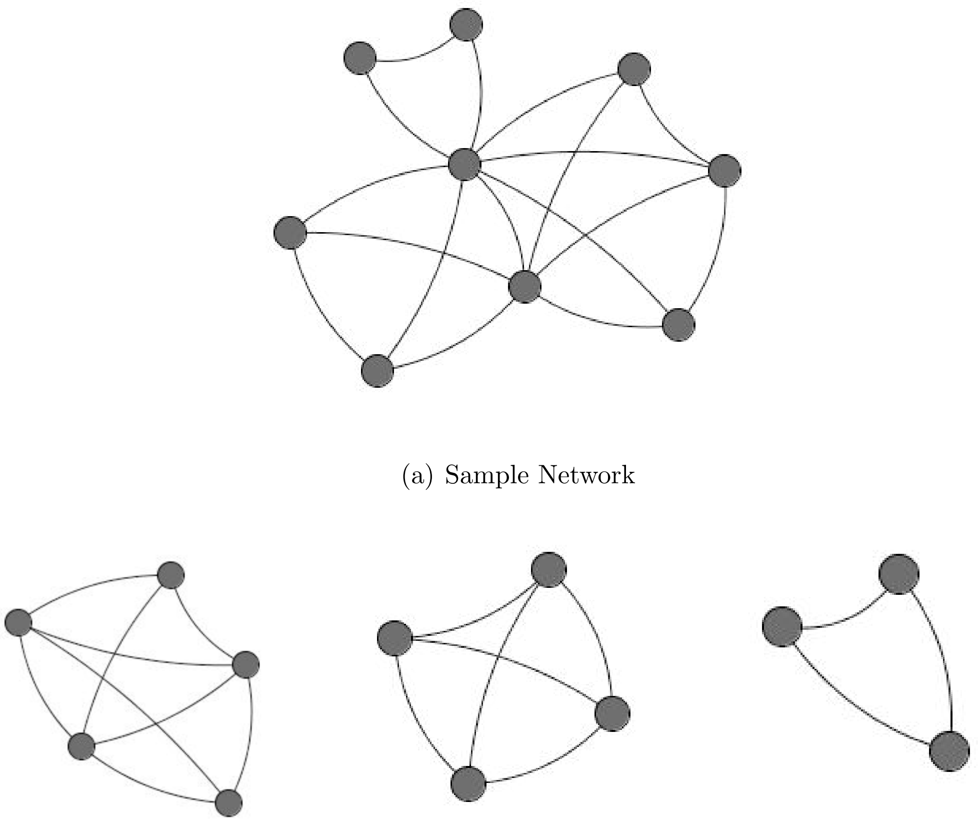
Faced with massive and evolving communication graphs, we use an approach relying on the unsupervised discovery of constraints based on content. By automatically extracting topic constraints from the message corpus a useful dissection of the massive communications graph into focused sub-graphs can be achieved.

Semantically Cohesive Networks

We believe the use of preprocessing or filtering to be superfluous when the analysis can yield previously unknown aspects about the massive graphs. We therefore argue that the utility of social network analysis can be enhanced when focusing the analysis upon a tightly defined and semantically connected sub-network within large diffuse communication networks. The analysis of this network will deliver a clearer understanding of the centrality and prestige in relation to a specific content focus.

The discovery of such semantically cohesive networks will be the starting point of our analysis (see section 2.1.2 on page 9). Since we will be focusing on the analysis of text message network corpora, we will consider sub-networks dedicated to a specific topic to be **semantically cohesive**.

Semantically
Cohesive
Networks



(a) Sample Network

(b) Sub-network A

(c) Sub-network B

(d) Sub-network C

Figure 3.2: Network Segmentation

3.4 Spectral Analysis

A simple yet effective approach mentioned in the beginning of this chapter is the use of content to partition the network into semantically cohesive networks, analogous to the separation of light of different frequencies when traveling through a prism. The dissection of the network into different sub-networks concerned with different topics can be described as a spectral analysis of messaging corpora.

The technique is illustrated in Figure 3.1 on page 38 in which the prepared massive network yields a typically constrained sub-network which in turn is treated to social network analysis. This approach relies on the generation of appropriate constraints which are used to extract relevant sub-networks. By using the content of the messaging corpus to find topics of interest we can postulate just such constraints[115, 82]. Each network by a topic can then be treated individually and with greater meaning than the graph as a whole.

An example consisting of a small synthetic network will illustrate the approach generally. Let us consider a small constructed network depicted in Figure 3.2. It consists of a small group of actors, of which the centrality of individual actors is intuitively observable.

Performing topic detection might then yield a segmentation of the network into three sub-networks, Figure 3.2(b) delineates a large sub-net concerning about half of the actors concerned, but semantically differentiated from the other two sub-networks depicted in Figures 3.2(c) and 3.2(d).

The approach utilizes several stages in which first of all topic detection must be performed before the subsequently filtered sub-networks can be analyzed. These stages will now be covered.

3.4.1 Topic Discovery

To prepare the groundwork for the segmentation stage we must first generate the constraint to be applied to the massive graph. In order to discover topic based sub-networks within the data we performed three basic steps:

Keyword extraction provides the initial keyword vectors per message. By breaking down the messages into keywords with the use of stemming and natural language processing techniques, we gain keyword vectors capturing the content of messages. These keyword vectors can be normalized using TF/IDF [4], and keywords which are either too frequent or infrequent are pruned. As is stated in chapter 4, we have used pruning to discard all keywords in the top 5 percentile indicating terms too frequent to carry specific meaning, as well the bottom 8% of terms occurring too infrequently to carry decisive meaning.

Network construction is the straightforward procedure of collecting all messages still left, i.e. not lost to pruning effects in the previous step. The graph constructed now is a sub-graph of the raw data referencing messages deemed relevant and of material impact to the analysis.

Topic detection can now be seen as locating groups of keyword vectors occupying similar locations in this vector space. Unfortunately, this high-dimensional vector space is unsuited to discovering such message clusters. We will therefore rely on singular value decomposition [55] to reduce the dimensionality of the keyword vector space used to capture the content of communication networks. Once a reasonably equivalent set of vectors residing in vector space of manageable dimensionality has been generated, groups of similar messages can be found. A common approach to finding such clusters has been the use of density based clustering methods. In this paper we will be using the DBScan [42] clustering method in chapter 4 to extract topics from messaging corpora.

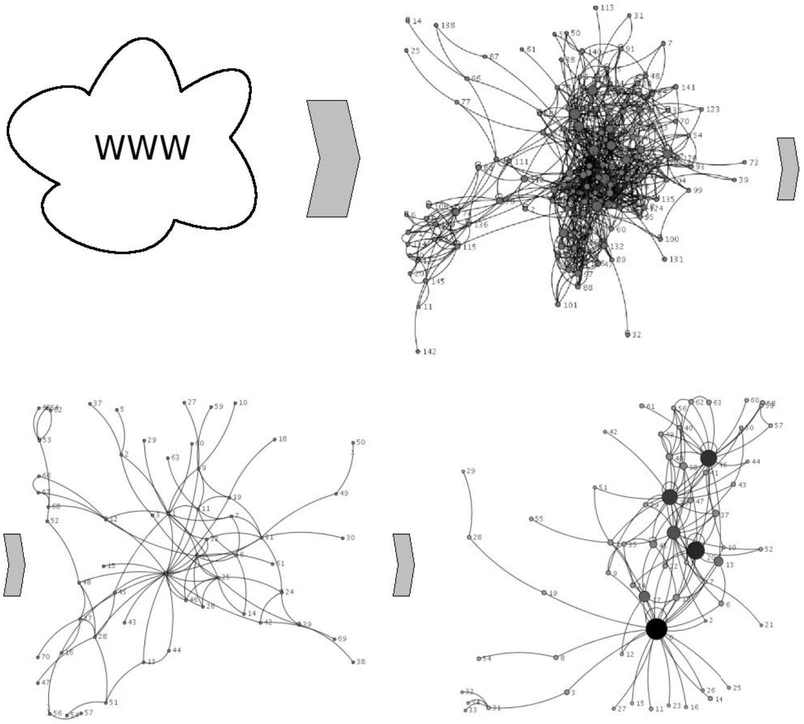


Figure 3.3: Segmentation Process

3.4.2 Network Segmentation

Using topics as a selection criterion, we can now associate messages with topics. Each topic is expressed as a sub-set of communications within the graph as a whole. By combining messages related to a specific topic, the messages for each topic cluster into a sub-graph of messages exchanged by the actors involved in this topic. There now remains a decision to be made in what fashion a topic related network is constructed:

Exclusive: The exclusive inclusion of messages in the topic cluster. Yields lower density sub-networks with a greater number of disjoint networks.

Inclusive: The inclusion of all messages of the actors mentioned in the topic cluster. Yields a denser more connected graph, but dilutes the semantic homogeneity.

A criterion for further analysis is the completeness of the sub-networks, it expresses the concentration of the topic within the network. Many small disconnected networks form a less satisfactory basis for social network analysis than a large complete sub-graph.

The sub-networks are defined by the semantic coherence, which we observe to coincide with topics. In essence, we note that any calculation of social network analysis measures over sub-networks of topics will yield applicable results, as the intention of the measures to be used on constrained communication networks is now given.

3.5 Content Based SNA

Using such specific network segmentation techniques raises the question of how such social network analysis based on semantically cohesive sub-graphs should be treated. This section looks at the issue of determining the centrality of actors with respect to a specific topic, and the relationship between this value and their overall centrality value.

These network segments can now be treated to social network analysis individually. Each sub-network is now able to determine the centrality and prestige of actors with respect to its own specific topic, which we refer to as **Topic Centrality** and respectively **Topic Prestige**. We find in chapter 4 that the relative centrality of actors is far more concentrated within sub-networks, and consequently bears a higher expressiveness.

Figure 3.3 shows a massive graph G being treated to topic extraction. The first network is the graph as a whole when extracted from the communications corpus. The second, and smaller network shows the sub-graph G' extracted with the help of topic T. This semantic cohesive graph can then be treated to content based social network analysis shown in the last graph.

3.5.1 Topic Centrality

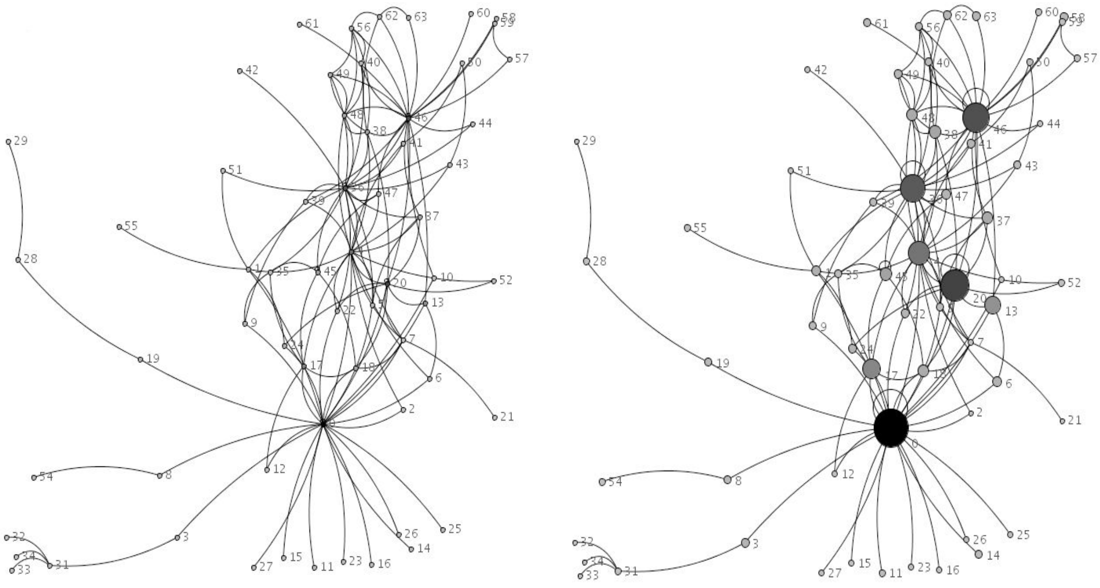
Topic centrality is the measure we arrive at when performing social network analysis on semantically cohesive sub-networks. This measure directly addresses the main problem raised when examining large networks, namely the loss of expressiveness of social network analysis.

To illustrate, a sub-graph extracted from a massive graph can be seen in Figure 3.4(a). The same network conveying emphasis of central actors can be seen in the adjoining Figure 3.4(b). Chapter 4 gives various examples of the segmentation process as well as the importance of this analysis.

This illustrates the benefit of using focused social network analysis constrained by semantic coherence.

3.5.2 Topic Prestige

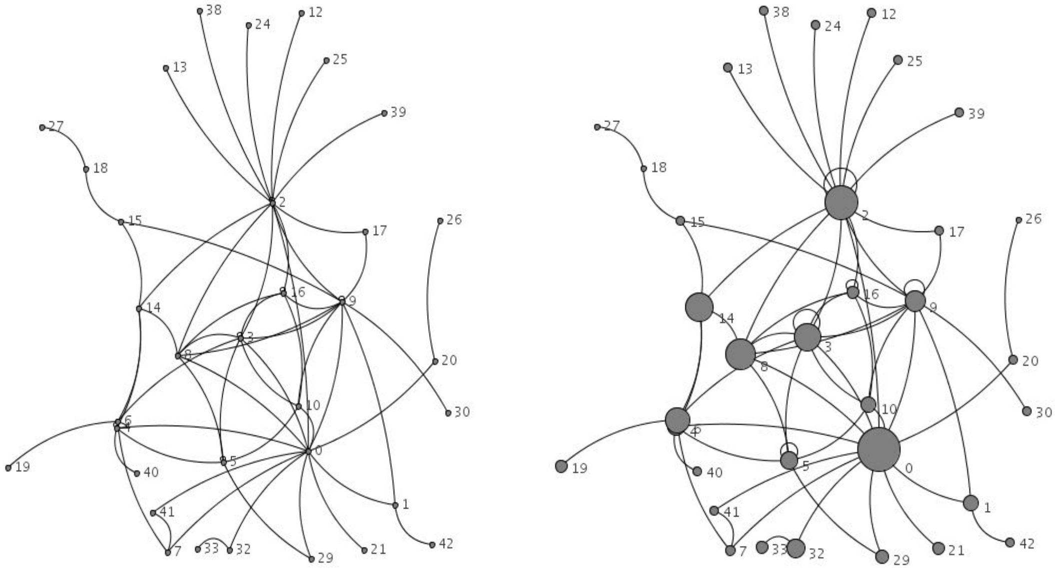
Analogous to the previous section, the relationship of an actors betweenness and their topic specific betweenness is scrutinized. Again the extracted sub-graph of a massive communications graph can be seen in Figure 3.5(a) and the application of social network analysis highlighting prestigious actors in the adjoining figure of 3.5(b).



(a) Topic Network

(b) With Emphasis

Figure 3.4: Topic Centrality



(a) An extracted topic sub-graph

(b) With prestige

Figure 3.5: Topic Prestige

3.5.3 Moving between topics

By definition a massive communications graph will consist of multiple layers of content, each topic ideally clearly defined. It is quite possible to observe an actor who is partaking

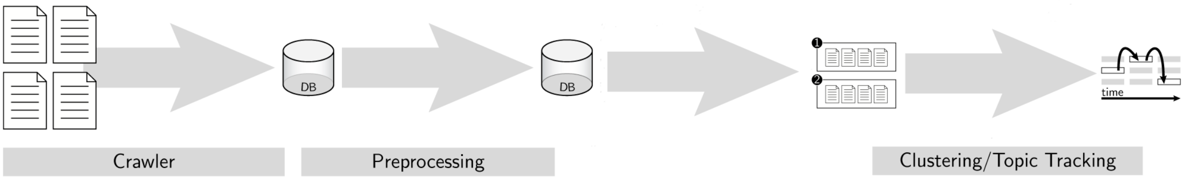


Figure 3.6: Tracking Topic Evolution

in communications within multiple topic contexts. Such an actor functions as a bridge between disparate semantically cohesive sub-networks, and as such forms an important intermediary junction across domains.

Actors playing a role in several topic networks can be seen to posses a form of centrality by merging the two networks. For instance, an Actor a in topic network T_1 is also present in topic network T_2. Cross topic centrality and prestige can now be calculated by merging the two networks T_1 and T_2 into a common network $T_C = T_1 \cup T_2$. Based on this common network T_C the centrality reflects the importance and relevance of a over both networks, ergo across topics T_1 and T_2.

The observation and characterization of such Cross topic considerations is a natural step when exploring complex communication networks. The exploration of this aspect forms an aspect of our future work.

3.6 Temporal SNA

Another aspect scrutinized is the development and evolution of topic semantically cohesive networks over time. To this end content spectroscopy is performed over multiple time frames to extract sub-graphs within designated time frames. These sub-graphs can then be analysed over time to follow the life cycle of such topics.

Figure 3.6 gives an overview of the proposed process. After the data has been gathered and treated to keyword extraction, there are several preparatory steps needed before topic evolution tracking can take place. This section shows this preparation, whereas section 3.6.2 will then go into the details on topic evolution.

The following sections will go into the details of the creation of a temporal topic hierarchy. Figure 3.6 shows the individual steps necessary for a single foreground model to be prepared, always keeping in mind that each foreground model calculated gives additional information based upon a unique time frame. A foreground model describes the word occurrences within a distinct time frame. The different views in conjunction, as shown in the general process graph in Figure 3.6 comprise the final temporal topic hierarchy. But first we will introduce the use of several time frames to capture varying topic life-cycle lengths.

3.6.1 Topic Structure

As delineated in section 2.3 on page 20, we commence the processing steps with the data in keyword/message format. We utilized the mentioned techniques of stemming, TF/IDF and removal of too frequent or infrequent terms to prepare the data. The implications of keyword extraction, stemming and related techniques are all preliminary to this approach.

The use of keyword vectors per message allows the use of clustering methods to discover clusters of semantically related content, which is the common approach to discovering topics in message corpora through the use of unsupervised clustering algorithms.

Data Segmentation

In order to be able to view topics persisting for varying lengths of time, we will use multiple sliding windows to segment the data. In this approach we have used two distinct time scales, one to detect long term trends and one to detect short topic instances. Each time frame will have its own sliding window length, and is processed independently from the other. It is quite possible, and will be investigated as future work, to use more than two window sizes.

For instance, we can use a long time frame of a month or more to cover the goings on over an extended time frame, and use a short time frame of a week to focus on local topics cropping up for shorter topics. Figure 3.7 shows the division of the time spanned by the data corpus being divided by two sliding windows of different length.

The selected window size should correspond to the expected topic time frame. For example, the use of 7 day slices as a basis would look for weekly topics (FG_s in Figure 3.7), with multiples of this basic time-frame as the longer term view (FG_l). In terms of notation we will refer to the sliding windows as foreground models, or FG windows. To discern different levels a subscript will refer to the length of the window (short or long), and a superscript to the index of the window, should one be necessary. For example FG_s^3 would refer to the third window of the window segmentation using short time spans.

Topic Extraction

After the data has been divided into distinct segments, the content of these windows must be determined. Each segment FG_x^i will be considered separately. Topic detection is performed in order to quantify topical interest within the specified time frames. These topics can be readily extracted with the use of existing clustering techniques. The keyword vectors comprising individual time frames can be clustered using, for instance, spatial clustering techniques, such as the density based algorithm DBScan[42].

As density based clustering shows poor clustering behavior when confronted with high dimensional sparse data, as is the case with the keyword vectors used in this case, we do not find topic clusters directly in the vector space. We use singular value decomposition [55] to find a low dimensional space with equivalent topology, losing some precision, but gaining a cleaner and faster clustering.

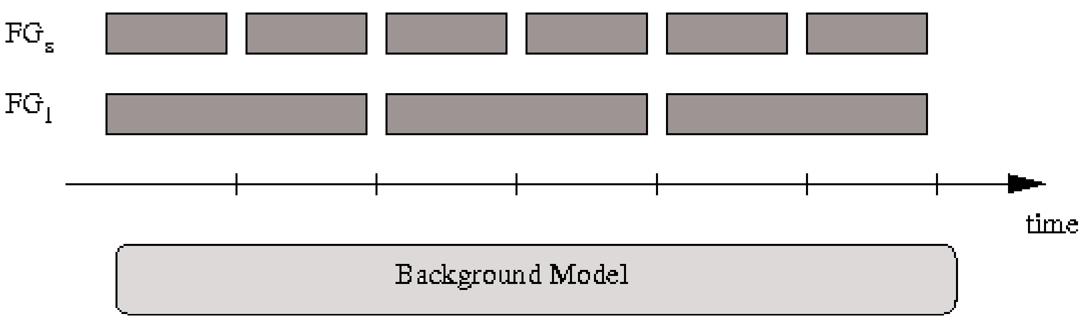


Figure 3.7: Fore- and Background Models.

Topic Keyword Discovery

It is important to note the fact that the approach proposed in this paper uses *multiple* time frames to capture topics using different time scales. In the following step we want to be able to quantify the similarity between topics of different time frames, i.e. long and short, in order to link trends to topics. We can use representative keyword vectors as a basis for comparison between FG_s and FG_l.

In order to extract representative keyword vectors binomial log-likelihood ratio tests to determine keywords important to a specific topic are used. In the following section we introduce representative topic vectors used to gauge the similarity (and hence relationship) between topics stemming from different time frames, i.e. foreground models.

In order to be able to compare topics, we designate the data corpus the **Background Model**. This language model is used to characterize the general communication being analyzed, thus allowing topics to be described by keywords which are more characteristic of the window the topic resides in with respect to the data corpus being analyzed.

The background model consists of the time interval denoted by the top bar in Figure 3.7. This need not be the entire text corpus, merely a significantly large portion to capture an adequate background. This can be said to be an extent covering most of the time period and should include a multiple of the largest time frame used in a foreground model. This view of the data captures all occurrences of keywords in the time interval, akin to the sum of all conversations in a room without differentiating between individual conversations.

The foreground models, used to gauge relevance of a keyword within a window, are denoted by the smaller time slices in the lower half of Figure 3.7. The two foreground models designated FG_s and FG_l reflect the use of multiple temporal views on the data. These views on the data capture conversations in a more focused setting, and will be contrasted to the background model to quantify the topical interest of the individual time slice.

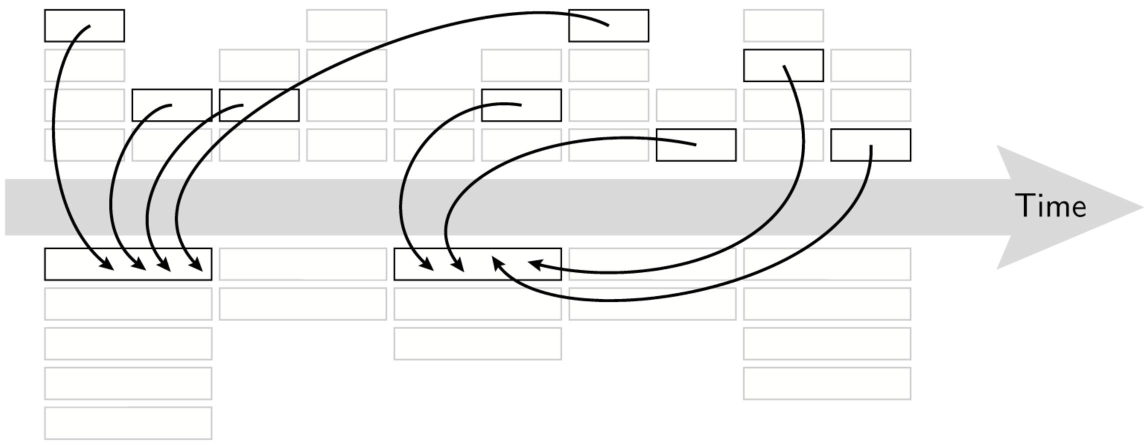


Figure 3.8: Modeling dependencies over time: Short term topics are linked to longer term trends.

Representative Vectors

With the help of the Binomial log-likelihood ratio tests (BLRT) [38] approach, the keywords engendered in a topic can be ranked according to relevance by contrasting the foreground with the background model. Each ranked topic vector can now be represented by an equivalent but shorter vector comprising only the most relevant contributors to the vector. In practical terms we have found relatively short vectors comprised at most of about 5 to 10 keywords to be quite sufficient.

BLRT forms the basis of our further considerations when attempting to find terms typical for a topic. This technique works well with both large and small text samples and allows for direct comparison of the significance of rare and common terms. The latter is particularly important as some probabilistic techniques produce unexpected results for infrequent attributes.

According to [38], about 20–30% of the vocabulary of typical English news-wire reports consist of words with a frequency of less than $\frac{1}{50000}$ in a moderately sized text corpus. Many of these rare words are content-bearing or technical jargon.

A word distribution that key topic terms should be identified in is derived from the topics found in the foreground model, the word distribution that is used as reference stems from the background model. For example, a foreground distribution of a topic consisting of news articles of the current week would be compared to a background distribution that consists of all news articles of the last six months.

This can be done by binomial log-likelihood ratio tests as they allow to compare two binomial processes. The following likelihood function has been used in our case: let p_i denote the probability of a word occurrence, n_i and k_i the occurrences of the words in the background corpus and foreground topic respectively. The index i indicates which two words are to be compared.

$$H(p_1, p_2; k_1, n_1, k_2, n_2) =$$
$$p_1^{k_1} (1 - p_1)^{n_1 - k_1} \binom{n_1}{k_1} p_2^{k_2} (1 - p_2)^{n_2 - k_2} \binom{n_2}{k_2}$$

For $p_i = \frac{k_i}{n_i}, p = \frac{k_1 + k_2}{n_1 + n_2}$, the likelihood ratio is calculated as

$$\lambda = \frac{\max_p H(p, p; k_1, n_1, k_2, n_2)}{\max_{p_1, p_2} H(p_1, p_2; k_1, n_1, k_2, n_2)} \tag{3.1}$$

The explicit form of this ratio can be rewritten as

$$\lambda = \frac{\max_p L(p, k_1, n_1)\, L(p, k_2, n_2)}{\max_{p_1, p_2} L(p_1, k_1, n_1)\, L(p_2, k_2, n_2)} \tag{3.2}$$

where $L(p, k, n) = p^k (1 - p)^{n-k} \binom{n}{k}$.

In order to find terms that are typical for a distribution, the likelihood ratio BLRT is calculated for each word of the topic vocabulary:

$$BLRT(w) = -2 \log \lambda = 2 \log \frac{L(p_1, k_1, n_1)\, L(p_2, k_2, n_2)}{L(p, k_1, n_1)\, L(p, k_2, n_2)}, \tag{3.3}$$

The parameters are set to the following values in order to calculate the applicability (or informativeness [112]) of each word:

$$\begin{array}{rcl} k_1 & = & C_{fg}(w), \\ n_1 & = & \sum_w C_{fg}(w) \\ k_2 & = & C_{bg}(w), \\ n_2 & = & \sum_w C_{bg}(w) \end{array} \tag{3.4}$$

$C_{fg}(w)$ and $C_{bg}(w)$ denote the frequency of w in the foreground topic with respect to the background corpus. The value $BLRT(w)$ denotes whether the probability distribution of a word in the foreground corpus is similar to the probability distribution in the background corpus. Word w_1 is differently weighted for the foreground model than word w_2 if $BLRT(w_1)$ does not agree with $BLRT(w_2)$.

Thus we can construct a keyword vector of terms relevant to the topic at hand, while at the same time being common to the data corpus at large.

By introducing two different time frames for the two models (one long-term, and one short-term), we are now able to gauge the applicability of terms in topics with their relevance in long term topic trends. We are also able to combine several short term topics into topic trend clusters. This view allows observation of a temporal hierarchy of topic life cycles within a general topic trend. The short-term topic tracking can be viewed as keeping tabs on local topic development, whereas the long-term view keeps the global topic evolution in perspective.

3.6.2 Evolution

It is the assumption of this approach that topics in news sources are covered at different (not necessarily consecutive) points in time. It is also assumed that salient topics are present for longer time intervals which allows us to use them as anchors in order to connect results of shorter time periods.

In order to be able to arrive at a temporal topic hierarchy, we at first need to detect topics at several different time scales. Each time scale presents a distinct view upon the data, determined by the length of windows used to divide the data into temporal segments for this scale. By introducing differing time frames when looking for topics, we are using topic discovery to lay the foundation of long-term hierarchical topic tracking.

We use the representative keyword vectors garnered from topics in the previous step to establish connections between distinct segmentation levels. This means short term topics are connected to long term trends, but neither are topics compared to one another nor are trends put into relation to each other.

Due to the curse of dimensionality, the comparison process yields significantly better results when distance comparison is performed in low dimensional space populated only by the most relevant keywords in contrast to the original high-dimensional feature space. Figure 3.8 shows the basic setup that is used.

Hierarchy Integration

We want to associate topics of a higher granularity with trends discovered in the long term segmentation. By considering topics found within the same window segmentation to be on one level, we connect topics from the short time-frame (FG_s) to topics found in the long term view (FG_l).

With the help of our previously (section 3.6.1) constructed representative vectors, we can now search for nearest neighbors among topics found by the foreground models. In our example we are looking at the two foreground models, the top half corresponds to a short time frame (such as the model FG_s^i in section 3.6.1), and a complementary long time frame (model FG_l^j).

The association of short-term topics with long-term trends is achieved through the use of distance measures. For the implementation presented in chapter 4 we have chosen the **Kullback-Leibler Distance** [79], thus using the natural distance between keyword frequency distributions between topics. By associating short-term topics with long-term trends we can build a two-tier temporal hierarchy.

Each longer term time slice can be associated with various short term time slices. The distribution and location of the short term slices can then be seen as instances when a long term topic crops up at different times and places as short term topics. Thus several topics from segments FG_s^1, FG_s^2, FG_s^3 and FG_s^7 are linked to a single long term topic from segment FG_l^1 in Figure 3.8.

Topic Evolution

By observing the topic instances of a given trend, we can observe how strong a trend is in different time periods. Densely populated topic clusters indicate a hot trend, whereas conversely more quiet news clusters show a more sedate trend. The time span covered by a trend, as indicated by the number and position of topic clusters linked to it, are also a good indicator of the persistence and hence importance of a trend. Thus a long term trend can be followed and gauged by the evolution of the topic clusters expressing this trend.

The use of a temporal topic hierarchy allows for a greater depth in understanding of topic evolution over time. We have worked in a two tier hierarchy for the exploratory work in this paper, but intend to investigate the impact of a greater number of hierarchy levels on content understanding and performance. The implementation of this approach can be found in the next section.

3.7 Summary and Outlook

This chapter provides the groundwork to the further exploration of semantically cohesive subgroups within massive communication graphs. By noting the decrease of relevance of social network analysis when analyzing increasingly large and heterogeneous communication graphs we propose to restore expressiveness to social network analysis by limiting the scope to sub-graphs with common topics.

By first treating a communications corpus to semantic analysis yielding discrete topics, we are able to focus social network analysis into a more viable and expressive tool when performing network analysis and actor characterization.

The use of semantically cohesive sub-networks allows the targeted analysis and tracking of topics within large communication corpora.

3.7.1 Future Work

The idea of topic centrality and topic prestige we propose in this chapter can go a step further in order to analyze the implications of social network analysis when actors have centrality values in multiple topic networks. In this case an actor can act as a connector between various topics, by constituting a knowledge or information bridge between discrete topics and their associated actors.

3.7.2 Density Based Sub-Graph Extraction

A recurring interest during mining massive graphs is the definition of interesting and valid sub-graphs upon which further analysis can be done. The use of subgroups is critical in such instances where density and connectedness form the basis of results yet also rely on the semantic consistency of the data such as in Social Network Analysis.

While the clustering of similar sub-graphs and frequently occurring sub-graphs has received a good deal of attention, the unsupervised extraction of relevant sub-graphs has received little or no notice. We therefore propose a mining approach for the unsupervised extraction of dense sub-graphs from massive graphs.

Discovery and analysis of sub-networks has been the focus of much research by bringing social network analysis into contact with graph mining. The extraction and analysis of user groups in communication networks [66, 45, 70] bring the tools of social network analysis to bear on groups within the daily communication flood.

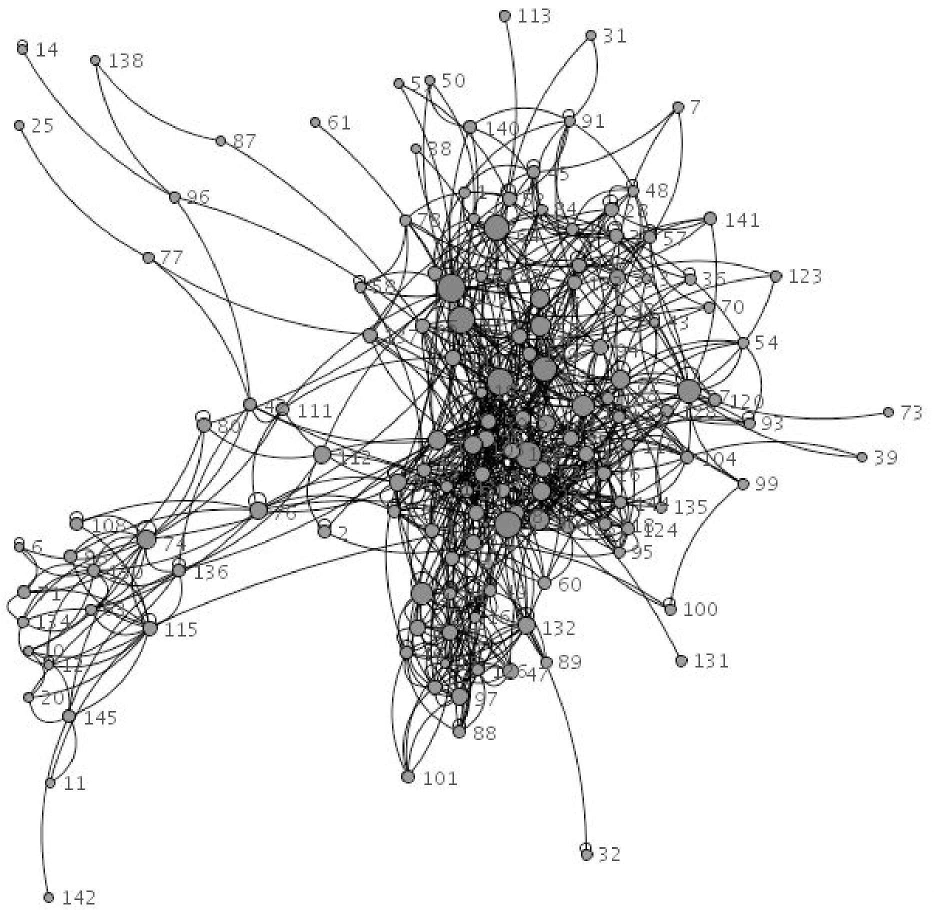
Graph mining has focused mainly on problems concerning frequent pattern mining[120], and the analysis and generation of power law graphs[24] has been well documented. An avenue of research has pursued kernel based clustering of graphs [73, 35].

We have looked at the possibility of treating graphs to density based clustering algorithms. As [35] notes, the applicability of kernel based clustering can be extended to the k-means approach. More generally speaking, we have noted the applicability of any density based clustering algorithm with the only requirement being that it must not rely on geometric information. Any comparisons between nodes of a graph must be done using a node distance based metric. The K-Median[119] approach satisfies this requirement, whereas the K-Means variant does not. We have modified two density based clustering algorithms to detect dense sub-graphs within massive graphs.

4 Case Studies

Contents

After the introduction to partitioning massive communication graphs in order to extract sub-graphs more amenable to further social network analysis in the previous chapter, we will now look at the implications and implementations of the approaches. We will walk through the partitioning process with respect to several communication graphs, e-mail data as well as blog data, in sections 4.1 and 4.2. Section 4.2.2 will also attempt to analyze blog data to find central actors and their evolution over time using the Eigenvalue approach.



(a) Communications graph

Figure 4.1: The Enron Network

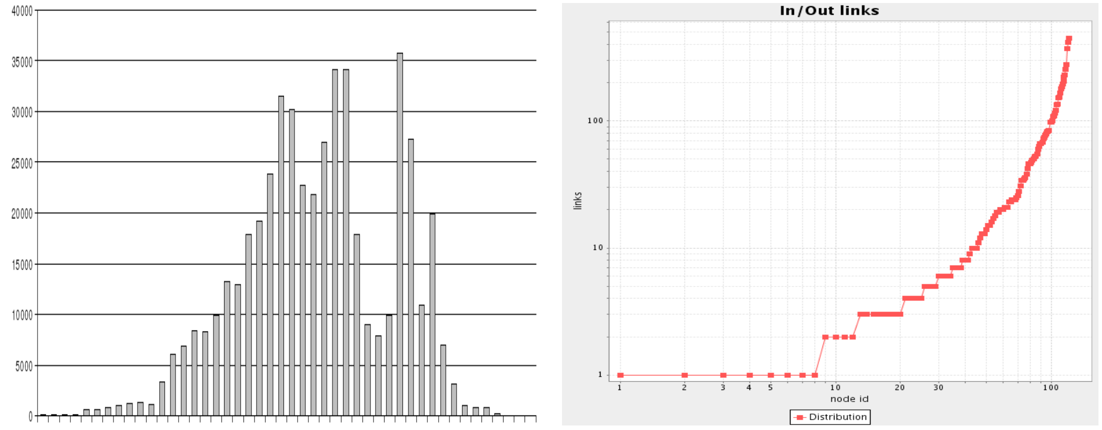
4.1 Partitioning E-Mail Environments

The Enron data set [30] is a portion of the emails collected by the prosecution during the Enron trial. The emails contained in this set have been moved into the public domain making this set one of the largest corpora of actual email data available today.

The main problem with the Enron data set is the fact that it represents a subset of a complete social network. The method of construction for this data set is described in the following pages, and the limitations discussed in the context of a communications network. The data set is comprised of 54 in-boxes, of which two actors are synonymous.

The Enron Data Set

The first data set under scrutiny will be the **Enron Data Set** [30]. The email network is comprised of 491.791 messages spanning the years 1997 to 2002. The number of



(a) Email activity from January 1997 to December 2002.

(b) Link distribution of the Enron data.

Figure 4.2: Enron corpus activity and corpus description

individuals partaking is around 42.000, an exact number is not possible to gauge as aliases crop up frequently, but not all can be recognized as such. The level of activity per month is captured in the 2 year time period between the middle of 2000 to the middle of 2002.

A caveat for this data set exists, as the raw data, comprised of emails recovered from servers from the time period of the investigation into the collapse of Enron, does not constitute a complete nor even comprehensive view of all communications having taken place at this time. The average daily activity per year lies around 2 messages per day, which given the size of the organization is a strong indication of the incompleteness of the data set. The nature of the data, as well as the format given, leads us to conclude that the data set originated as the confiscated or reconstructed in-boxes of 150 Enron employees and board members.

4.1.1 Data Description

The raw data is parsed to extract information regarding origin, destination, time and content of the email. The communications network as described by these emails is then distilled into an **Adjacency Matrix** of actors describing the network for further analysis. A great number of actors found in this network are external to the company, and have only a negligible number of interactions between actors associated with the Enron corporation. For this reason we have decided to focus on the core network of e-mail communications flowing between actors associated with Enron.

The network is filtered to remove any emails not originating as well as terminating within the set of actors. The basis for further analysis now contains about 2500 messages between 145 actors. This reflects the fact that all data collected for this set stems from

Email Subject Line	Content Categorization
Fw: Jim Lehrer interviews Gore and Bush	Political Humour
Fwd: AND I THOUGHT I WAS HAVING A BAD DAY !!	Anecdotes
Fwd: Witch's Brew	Anecdote
FW: DARWIN AWARDS	Anecdotes
Fwd: hey, girlfriends!	Anecdotes
Hit the Floor - True Story	Anecdotes
FW: You didn't get this from me	Jocular humour
Fw: FW: Fw: The moves]	Jocular humour
NOTICE OF REVOCATION OF INDEPENDENCE	Political Humour
RE: brain food	Anecdotes
Fw: Stupid is as Stupid Does	Anecdotes
FW: GAME DAY...could be one of the funniest emails ever. especial	Anecdotes

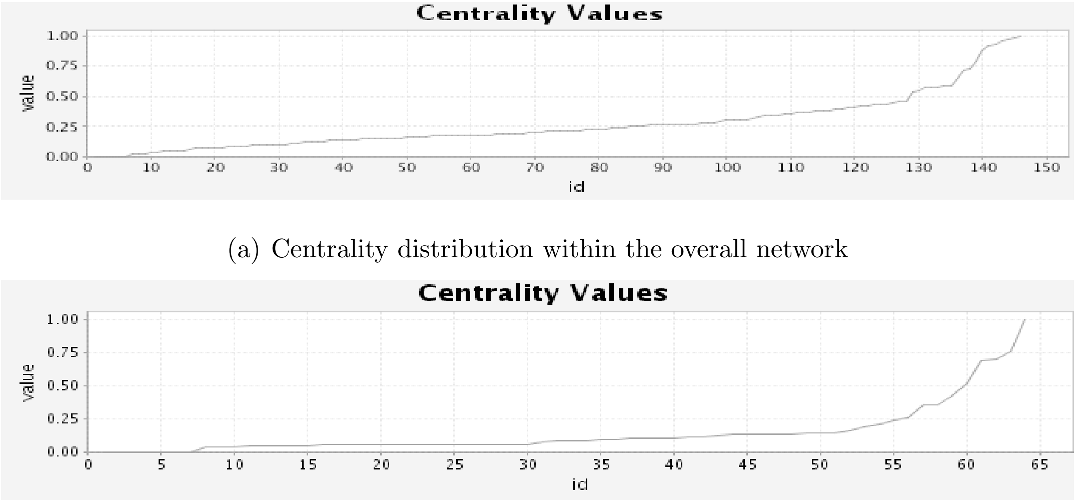
Table 4.1: Sample Cluster Results from the Humour Topic

147 separate in-boxes, two of which are aliases. This set of messages now constitutes communications purely internal to the organization, and can as such be seen as the corpus of ”private” data.

The resulting communications network can be seen in Figure 4.1(a). A visual scan will indicate one large strongly connected sub-graph with a few outliers. We have performed **stemming** and **keyword selection** processes to identify important words, identifying approx. 3000 keywords used in the prepared corpus. The application of singular value decomposition produces a reduced keyword matrix consisting of just over 200 dimensions covering 97% of the data.

4.1.2 Topic Network Extraction

Whereas the network structure is now adequately defined, the steps needed to perform topic detection rely on further analysis of the message content. We have performed **stemming** and **keyword selection** processes to identify important words, upon which similarity can be defined. Also the reduction of too frequent and infrequent words leads



(a) Centrality distribution within the overall network

(b) Distribution of centrality within the sub-network

Figure 4.3: Centrality distributions

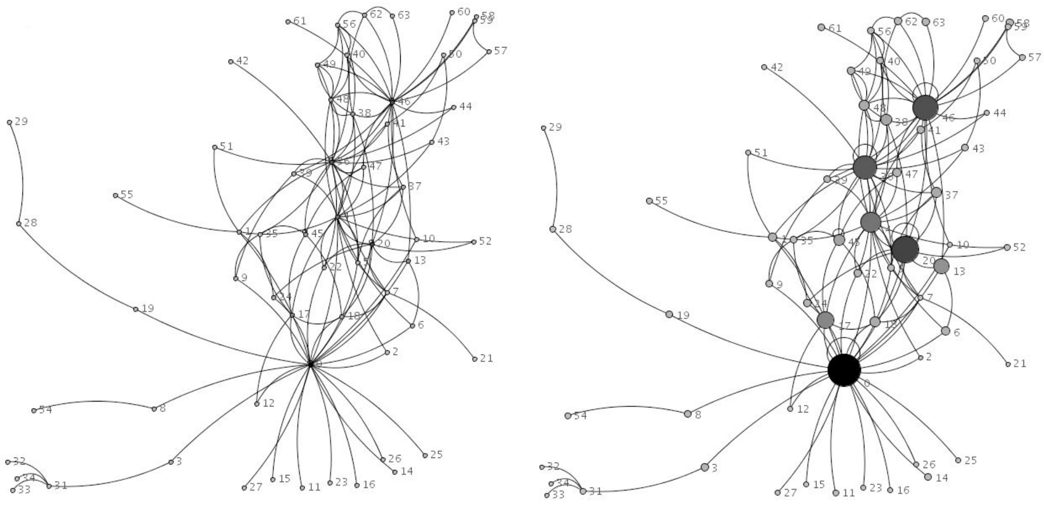
to the reduction of our feature space.

Clustering of the data has given us 20 clusters. Of these clusters we differentiate between a large noise cluster of almost half the messages. Of the remaining 19 we have a somewhat large cluster of 264 messages, with the remaining 18 clusters having only between 5 to 35 messages.

The small clusters are very tightly defined, the content of which can best be described as a thread of constant quoted replies where some messages are missing. We believe the relevance of these clusters suffers greatly from the fact that not the entire communications network, but that a rather arbitrary subsection is being mirrored in the data set.

It is interesting to note that the large cluster is indeed homogeneous enough to allow characterisation. This cluster contains many of the daily humorous and anecdotal emails being sent around offices today, and as such is an extremely pervasive and common topic found in most in-boxes. Almost half of all actors in the network under scrutiny have kept or archived such emails.

This cluster reflects the given nature of communications we intend to analyze, if not in seriousness then in ubiquity. Whereas a social network analysis of the entire time period will certainly identify the people central to the communications flow in the company, the selection of the humour sub-network can give us an insight as to who might be characterized as the greatest prankster. A small selection characterizing the office humour character is shown in Table 4.1.



(a) The humour sub-network (b) Central actors

Figure 4.4: The Enron humour sub-network

4.1.3 Topic Centrality

Performing social network analysis upon the whole network will of course establish a ranking on the relative importance of actors in a network based upon their propensity to talk to other people. In this case we find a high value of centrality for actors engaged in the management of the company, which is the expected result when the communications of a hierarchical organisation is analyzed. The higher the actor is located, the more he relies on communicating with others. This network of course contains all communications under scrutiny, thus limiting the expressiveness of the result to a statement of general importance. The centrality distribution for the entire network can be seen in Figure 4.3(a).

When considering the humour sub-network (depicted in Figure 4.4(a)), we can see that about half the actors within the entire network are to be found within this sub-network. Applying social network analysis to this sub-network (Figure 4.4(b)) shows us the existence of a handful of central actors, who apparently enjoy sending around office humour.

The centrality of these actors (Figure 4.3(b)) is far more pronounced with respect to the overall network (seen in Figure 4.3(a)), allowing for a more concise appreciation of the actors involved. The entire network has a sizable subset consisting of 17 actors with a normalized centrality value larger than 0.5, giving a diffuse picture of importance within the network. In the humour sub-network there is a better characterisation of central actors, as only 5 score better than 0.5.

In our approach we have combined the use of content analysis to constrict the communications network to be analyzed with social network analysis tools. The sample topic network we are looking at yields centrality results displayed in Table 4.3. As we

Actor	Centrality	Position
Andy Zipper	0.088335	Vice President Enron Online
Louise Kitchen	0.078280	Founder EnronOnline
John Lavorato	0.076380	CEO of Enron
Barry Tycholiz	0.071972	Vice President
Tana Jones	0.055695	N/A
Mike Grigsby	0.047822	director of corporate strategy and development.
Richard Sanders	0.047530	Vice President and Assistant. General Counsel for Enron Wholesale Services,
Geoff Storey	0.045056	N/A
David Delainey	0.042674	CEO, Enron Energy Services.
Michelle Cash	0.037957	Assistant

Table 4.2: Top ten centrality results over the entire network

can see, the top ten people with respect to topic centrality are now less represented by actors located higher in the corporate hierarchy. This reflects the nature of the topic in question, as we all assume that office jokes are not generally spread by managers.

Succinctly put, by performing topic discovery on the Enron data corpus, we can ask more specific questions about who does what. The use of topics to focus social network analysis increases the expressiveness of centrality and prestige when compared to the complete network.

4.2 Web-Log Environments

In order to ascertain the applicability of the proposed approach from section 4.3, we have selected two text based messaging corpora. Both corpora are electronic communication networks, the first a weblog capturing conversations in post and response form, and the second an e-mail network. By applying **content-based network segmentation** to these networks we will demonstrate the capability to extract actors of importance to specific domain knowledge without apriori knowledge about structure or content of the text-based messaging corpora.

Actor	Centrality	Position
Richard Sanders	0.048980	Vice President and Assistant. General Counsel for Enron Wholesale Services
Susan Scott	0.042857	N/A
Jeff Dasovich	0.036735	N/A
Elizabeth Sager	0.027959	Legal Department
Mark Haedicke	0.027755	Attorney with Enron North America
David Delainey	0.027347	CEO ENRON
Tana Jones	0.025102	N/A
D. Steffes	0.018367	N/A
Richard Shapiro	0.016735	Vice-President of Regulatory Affairs
John Lavorato	0.015918	CEO of Enron America

Table 4.3: Top ten centrality results table over the topic network

We have implemented the introduced approach in order to verify the existence and relevance of topic based sub-networks in real world data. The approach introduced in Section 4.3 has been applied to the Enron network, as well as to a blogging network.

The use of social network analysis has found increasing application to communication structures of varied origins in recent years. The structure of newsgroups, web-logs, email networks and information dissemination are just a few of the data sources. It is this subset of data sources concerned with messaging networks which we will now focus upon.

The use of news groups and similar messaging boards has provided a bountiful source of data for the inquiry into the structure of topics during discussions. One central aspect when such data sources are concerned is the interpretability of the centrality and prestige in these networks. More than in other areas, the precise meaning of social network analysis is very much dependant upon the selection and cleaning of the data in preparations to analysis. For instance, the size of a network has a direct impact on the relative achieved scores, as it is harder to achieve high centrality in large networks.

In order to ascertain the applicability of the proposed approach, we have selected two text based messaging corpora. Both corpora are electronic communication networks, the first a weblog capturing conversations in post and response form, and the second an e-mail network. By applying **content-based network segmentation** to these networks we will demonstrate the capability to extract actors of importance to specific domain knowledge without apriori knowledge about structure or content of the text-based messaging corpora.

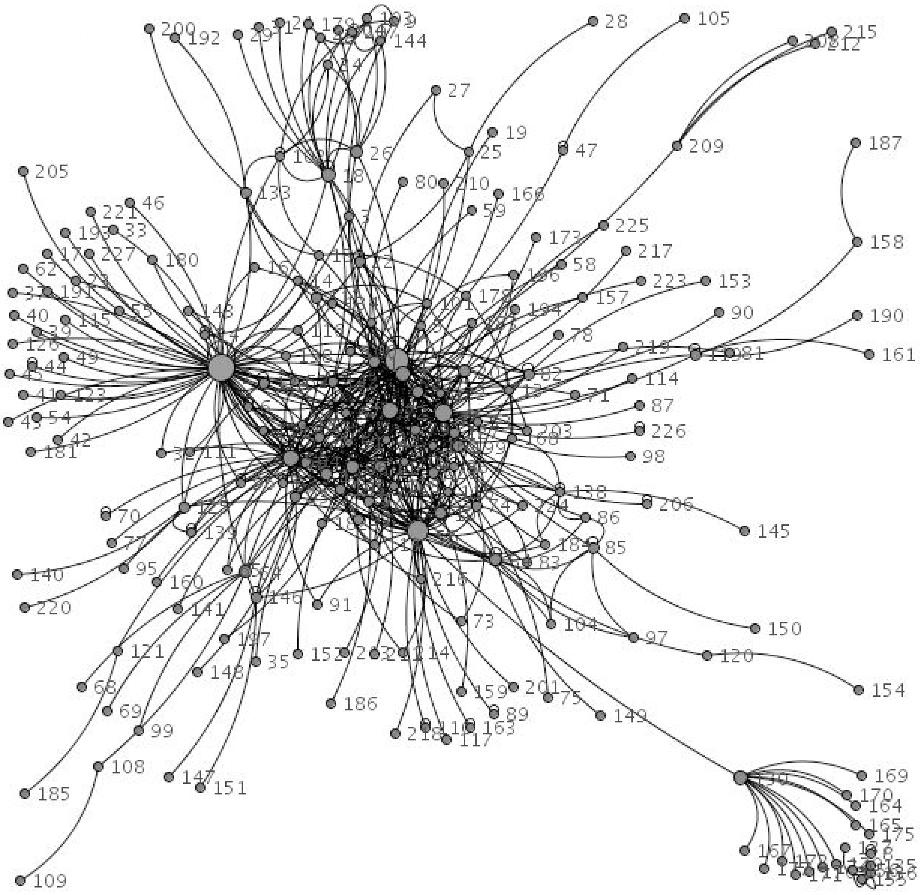


Figure 4.5: Weblog Communications Network

4.2.1 Partitioning a Blogging Network

The first data corpus we will be analyzing is a blogging communications network gathered from the intranet of a large corporate environment. The corpus consists of 438 actors communicating a total of 5732 times over a period of six and a half months. The shape of the network can be seen in Figure 4.5.

Data Description

The stages concerning data cleaning and preparation produces a keyword matrix of approximately 2300 dimensions, which after reduction of the too frequent and too infrequent terms is reduced to about 1700 dimensions. By performing singular value decomposition upon this sparse matrix, we were further able to approximate the keyword matrix with a reduced matrix consisting of just less than 150 dimensions covering about 95% of the data. This matrix forms the basis for topic detection.

Cluster #	Size	Topic
1	839	Running a Blog
2	592	Soccer, World Cup Soccer
3	381	Exchange of working and living experiences from different cultures

Table 4.4: Topics of the blogging network

Topic Network Extraction

By performing density-based clustering using the DBScan algorithm we were able to extract 17 different topic clusters from the data corpus. These clusters make up just over 60% of the data, with the rest being classified as noise. The three major clusters making up the bulk of the clusters talk about the creation and running of a weblog (832 messages), soccer and the world cup (592 messages) and the comparison of living and working practices in various countries around the world (381 messages). All three clusters are well defined, having more than three quarters of the messages on topic. The remaining clusters are less meaningful, and are all smaller than about 100 messages each.

Topic Centrality

By focusing on the network discussing blogging, setting up a weblog, and etiquette during blogging, we are able to extract an aspect of the network hidden in the data before. This sub-network has a strong central actor forming a focal point for the entire community. In contrast the centrality values for the entire network which are solely reminiscent of the scale-free nature of the network, with the centrality corresponding to verbosity. Figure 4.7(a) shows the extracted topic network.

The normalized centrality of actors is displayed in Figure 4.7(c). We can see the pronounced reduction of important actors in such a cohesive sub-network in comparison to the entire network, whose normalized centrality values are seen in Figure 4.6. The

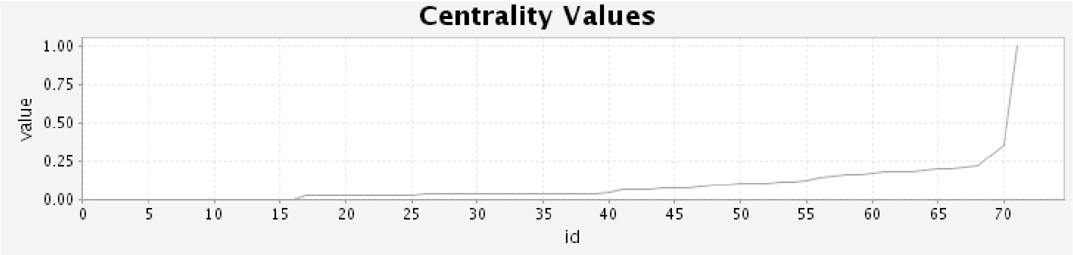
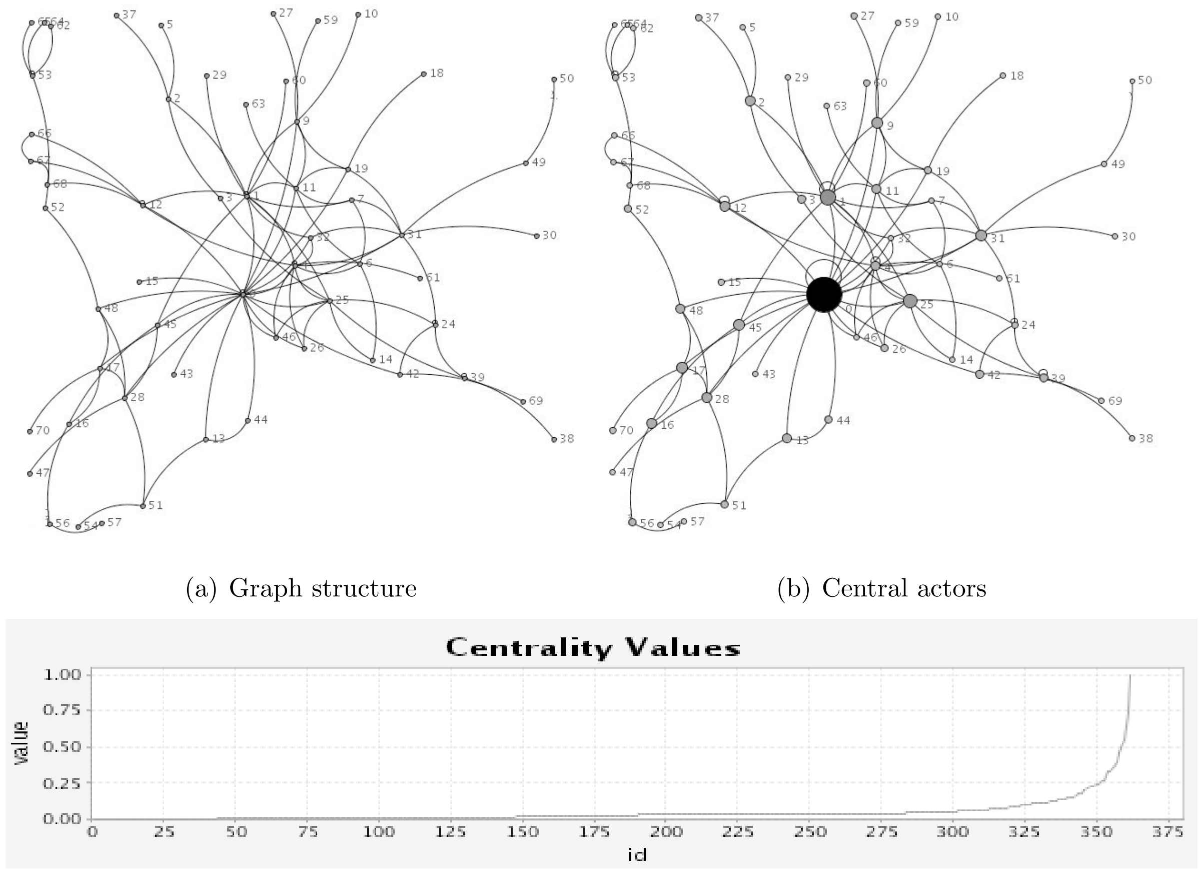


Figure 4.6: Distribution of centrality within the weblog network



(a) Graph structure

(b) Central actors

(c) Distribution of centrality

Figure 4.7: A topic within the Blogging Network

central actor of the sub-network scores a (non-normalized) value of 0.57 within the sub-network, but an undistinguished 0.0335 within the entire network. Figures 4.6 and 4.7(c) shows the improvement of the numerical centrality values, and the increased usability of the topical centrality makes these values much more expressive. Figure 4.7(a) shows the extracted topic network, and Figure 4.7(b) highlights the importance of an actor in the center of the network. This central actor represents a significant improvement in understanding of the data, as an actor with previously unknown domain knowledge is found, who is considered a touchstone of the community.

Figure 4.7(b) highlights the central actor of the cohesive sub-network. This actor represents a significant improvement in understanding of the data, as an actor with previously unknown domain knowledge is found who can be considered a touchstone of the community.

4.2.2 Eigenwert SNA

We also used social network analysis approach for trend and trend-shift detection over time in a technically oriented newsgroup on mobile phones. The analysis was based on the assumption that a trend shift is relevant only if relevant (central) members of the newsgroup initiated this shift. A shift could occur in one of two ways. Either within a subgroup the topic shifted, e.g. other words were used, or the relevance of a group within the newsgroup shifted, and the now more relevant subgroup discussed a different topic. We used eigensystem analysis as a method and could show that as groups shifted over time, so did topics.

Data

In a world of short product life cycles, especially in the telecommunication and mobile phone markets, anticipation of trends is of utmost importance for companies that want to survive in this highly contested markets. The anticipation of trends has been based on marketing research efforts alone in the past. In the age of the Internet users share their information, perceptions and thoughts with other like-minded users for example in newsgroups. Methods based on social network analysis can be used to evaluate this freely available and information rich communication to observe what users feel and think about products, techniques and or services. Forecasting trends from these communications and developing the right products and services at the right time could bring a crucial advantage to any company capable of performing this kind of analysis.

To achieve this goal, not only does one have to take a look at what users write about, but who is writing. A topic may not be worthwhile if the author has no influence within the group. But if a highly esteemed author in a group brings up a new topic he claims as relevant it may be of relevance to the investigating company. Thus it does not suffice to look at words or phrases used in the context of the newsgroup, but one also needs to incorporate the knowledge about the social network and its structure underlying this newsgroup.

In this project a joint analysis of the social network and the topic structure was developed and tested. The objective of Corporate Technology was to find a way to incorporate the knowledge of the social structure of a group of mobile phone users with the knowledge about the topics about which they communicate. The data set which was collected and preprocessed by two of the authors is from a newsgroup on mobile phones. The preprocessed data set was imported into a SQL data base specifically adapted for the further analysis of this study. Overall the data set consists of 2808 messages written by 709 authors during the observation period from July 2005 to October 2005.

Based on that standard approach we use the complex eigenvector centrality as defined by [68]. It is based on a complex hermitian adjacency matrix H that is derived from the real valued weighted adjacency matrix A of the underlying network by Eq.(1).

$$H = (A + iA^t)e^{-\frac{i\pi}{4}} \tag{4.1}$$

where $i^2 = -1$ is the imaginary unit and A^t is the transpose of A. Thus directional information about the communication is kept. Hermitian matrices have a complete orthogonal eigensystem, all eigenvalues are real, the eigenvector components may be complex valued. The set of all eigenvalues is termed the spectrum.

The complex valued eigenvector components are interpreted as a rank prestige index of each subgroup member, following the interpretation of the real valued eigenvector centrality index described for example in [121]. In addition, each member has for each subgroup structure / eigenvector in the spectral representation a different rank prestige index. This index depends on his relation to the respective anchor of the subgroup. The eigenvectors themselves represent different patterns of communication.

The interpretation of a complex eigensystem can be based on two extreme examples (based on an adjacency matrix with $h_{kk} = 0$). The first is a star graph. The spectrum is symmetric in the sense that the non-zero eigenvalues come in pairs of $+\lambda_k, -\lambda_k$, with k defining the number of star centers, and the corresponding eigenvectors show a phase shift of π between components that are in contact with each other, e.g. belong to the same star. The other extreme is the complete graph. Here the spectrum is organized in such a way that $\sum_{k=2}^{n} \lambda_k = \lambda_{max}$. The eigenvector component with the highest absolute value is regarded in this paper as the anchor or most influential node or person of a pattern. The eigenvector components within each eigenvector sorted by absolute value give the generalized index of centrality within the pattern described by the eigenvector. Thus a node with a relative high absolute value in a given eigenvector is regarded as more central than one with a relative low absolute value. The phase of each eigenvector component describes the direction of communication this node has within this pattern with respect to all other members.

Results

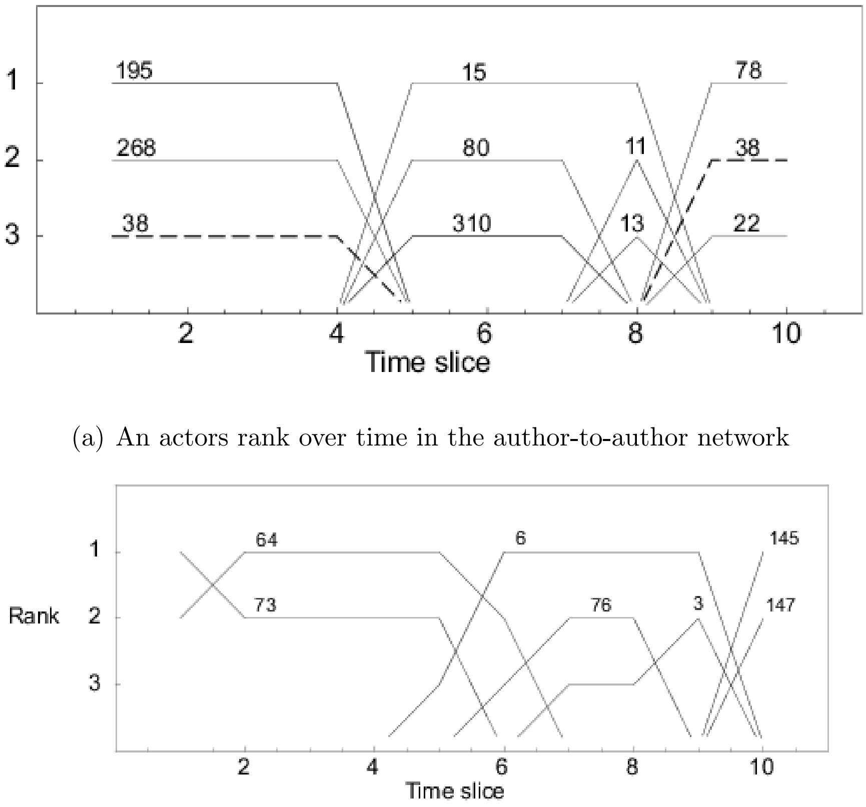
To achieve our goal we used two different adjacency matrices derived from the data set. First we created the author-to-author network, so as to find the relevant authors and their subgroups. This was achieved by setting a link between author k and author l if the message by author l was a reply to the message by author k. Thus we used the message identification in the mail header to link authors, because in a newsgroup the reply is not to an author but to a message. In the next step we analyzed a two-mode network consisting of a matrix composed of authors (rows) and the words (columns) they had used, e.g.:

$$C = \begin{pmatrix} 2 & 5 & 7 \\ 0 & 2 & 8 \end{pmatrix} \tag{4.2}$$

The adjacency matrix A was constructed as a row + column square matrix by:

$$A = \begin{pmatrix} 0 & C \\ 0 & 0 \end{pmatrix} \tag{4.3}$$

And A was transformed by Eq. (1) in to the complex hermitian adjacency matrix H.



(a) An actors rank over time in the author-to-author network

(b) Change of anchors over time for authors-use-of-words

Figure 4.8: Rank of actors in a blogging network

Combining the results of the eigensystem analysis performed on each matrices we found that this approach shows some potential to identify topic shifts.

To see the development of the subgroups and topics over time, we adopted a sliding-window-approach. A time slice holds the communication of 28 days. It is then shifted by 21 days. We decided on this approach so as to make sure that relevant shifts can be detected. If on the other hand one would use adjacent but non-overlapping time windows one might miss relevant events if they fall into a time single slice.

The eigensystems have been calculated using the function Eigensystem of Mathematica [129]. To ensure privacy for the authors we use IDs. Since the two networks are of different size, the IDs change. In table 4.2.2 on page 68 IDs of authors in the author-to-author setting and in the author-to-words setting are given to allow the reader to identify one author in both networks. Whenever both settings are compared we use a pair e.g. (15,6) to identify the author.

Author-to-author network

The author-to-author network was calculated on the complete data set. Figure 4.8(a) on the preceding page shows the shift s of the anchors of the first five eigenvectors over time. They have been ranked by the value of the corresponding eigenvalue. Thus the top line represents the most relevant author. Shifts occur in periods 4 - 6 and 8-10. The sliding window approach shows this development nicely. In detail:

- There is a shift from author 195 as the most prominent member of the newsgroup to author 15. This shift occurs in time slices 4, 5 and 6. In time slice 6 author 15 takes over completely. He 'won' against author 281 who was also close to gaining the group-anchor position.

- In time slice 8 already author 78 becomes visible. In time slice 9 we see how he starts to take over from author 15. In time slice 10 he takes over completely.

If we look at the evolution of subgroups it was found that the subgroup around author 195, consisting of authors 268, 154 keeps the central position in the first three time slices. In time slices 2 and 3 there is already a competition in eigenvectors 3 to 6 for the 'take over'. One group is author 193 together with author 543, the other is author 15 together with author 596. These two groups seem to be of almost equal traffic strength and communication behavior, but seem to exclude each other. In time slice 4 the group around author 281 takes the role of the group around author 193. Author 281 goes together with author 154 and thus links this subgroup with the prominent group around author 195. In time slice 5 author 195 seems to have given up his central role. The subgroup around author 281 seems to be a very stable and undisturbed subgroup with members 154, 528. In time slice 7 author 15 finally takes the central position with the subgroup consisting of authors 388 and 289. In time slice 8 author 78 starts to move forward. In time slice 9 he starts to take over the leading position from author 15. This subgroup consists of authors 38 and 22. In time slice 10 we see that author 78 has taken over.

Authors-to-words network

The analysis of the use of words by authors has been performed on a subset of the original data set. The matrix has been constructed as follows: First all authors and all words were used to create the matrix; the authors defined the rows and the words defined the columns of the $n \times m$ matrix; then a threshold was defined which was applied to the row sum and column sum. Thus only those authors were taken that had used in total more words than the threshold defined, and only those words were kept that had been used more often by authors than had been defined by the threshold. The threshold has been arbitrarily set at 12. This resulted in 144 authors and 864 words being used. In the following we will show the original words in inverted commas. Thus e.g. 'handi' is the word that was found in the messages. Please be aware that 'handi' is a word used

Author-to-author	Author-to-words
15	6
195	64
268	73
281	76
78	30
193	63

Table 4.5: IDs of the same author in different networks

in german for mobile phones. It sounds english, and thus should be written as 'handy', but this is not the case in the original data.

In Figure 4.8(b) on page 66 the time development of the author-to-words network can be seen. The anchors in most time slices (IDs (195,64), (268,73) and (15,6)) are exactly the same as in the author-to-author network.

In the first two eigenvectors of each time slice only the central authors and their most often used words are given, thus revealing no specifically relevant information. The 2nd and 3rd eigenvector show the split in usage of words and groups. Starting from the 5th and 6th eigenvectors it seems that more interesting words like 'icq' show up. In tables 4.5(a) on the next page and 4.5(b) on the facing page the results are listed. Only the most important authors and words in each substructure are shown. The construction of the two-mode matrix as shown above results in a graph with star like substructures. These lead to an eigensystem as described in section 2, where each star like substructure is explained by two eigenvectors. What can be seen here is that as groups shift, topics shift - albeit with minor lags.

The first two eigenvectors show author 64 as the anchor. Authors 73 and 6 show a very similar strength and use of the same words. But since the main word is 'handi', this is not a surprising result. It may be of interest that the word 'siemens' plays a visible role in the first two eigenvectors, while the word 'nokia' first starts to show up in the third and fourth eigenvector. The three authors are the most prominent as has already been shown in the overall and the time dependent analysis. The first two eigenvectors give the 'groundswell' of the traffic. To find more interesting results one should take a look at the following eigenvectors. Here we start to see splits between sub-patterns within one eigenvector.

In the third and fourth eigenvector there seems to be the following split: author 73 is the anchor with 64 as a follow up author. They use the word 'profil', while 6 and 76 seem to talk about 'nokia'.

When looking for words that might be interesting, the seventh and eighth eigenvector start to show the word 'icq' within a subgroup consisting of 6, 73. The other group

(a) Use of words in time slice 1 to 3

Eigenvector	Authors	Words
1,2	73, 64	'profil'
3, 4	64, 6 and 75, 73	'handi', 'profil', 'entfernen'
5,6	63, 48	'nokia', 'icq'

(b) Use of words in time slice 6

Eigenvector	Authors	Words
1,2	6, 64, 76	'handi'
3, 4	6, 64	'newsgroup'
5,6	63, 6, 64	'icq', 'improve', 'nvideo'

Table 4.6: Words used in time slices

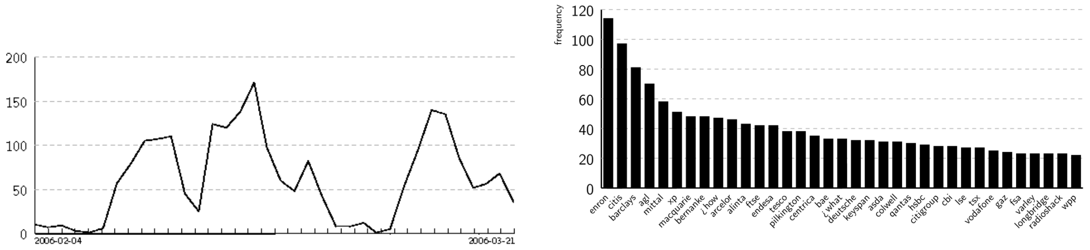
consists of 76, 64 and 48. They talk about 'newsgroup'. Of interest might again be the use of the word 'icq' by and 64 and others, and the words 'nokia', 'akku' and 'telefon' by a subgroup consisting of 6, 73 and others. The 18th and 19th eigenvector seem to have to do with Palm-technology and maybe the use of Outlook, since these are words used in these eigenvectors. 'motorola' also is used in that conjunction. The subgroup is centered around author with ID 3. Due to the directional information inherent in the complex valued hermitian adjacency matrix and the consequent eigensystem analysis a second relevant subgroup (authors 75 and 6) with the interesting word 'akku' was found.

The first major change seems to happen between time slices 5 and 6, the second shift occurs in time slices 8 to 10. This corresponds to the structure change in the author-to-author network. In Tab. 3 the changed groupings are given.

Conclusion

As a conclusion it can be stated that the method proposed can yield interesting insights into the 'mechanics' and 'dynamics' of communicating groups as could be shown in this feasibility study. To gain valid insights from the combination of the analysis of the communication behavior with content analysis it is necessary to perform a considerable effort in data preprocessing such as natural language processing within a concept tailored to the goals of the analysis and the linguistic specifics of the text corpus. Especially the dynamics of actor-to-actor communication and the dynamics of use-of-words make this necessary but worthwhile.

Furthermore, the data of this project suggests that changes in the social structure are leading indicators for topic changes. However, this has to be corroborated by further research. A comparison with other clustering methods has not been in the scope of this



(a) Observed Publishing Activity

(b) Word occurrence within the corpus

Figure 4.9: Blog activity and word complexity

project. However, a meta-study of Freeman [50] compares 19 methods for the analysis of a two-mode data set and finds that eigensystem analysis is one of the top-ranked methods.

4.3 Topic Tracking

We have evaluated the use of hierarchical temporal topic tracking on the basis of real world news data sources. News articles are expected to be a good quality data source for clustering as they are much longer than for instance newsgroup articles and are written by professional authors or derived from content that was bought from one of the large news agencies (such as Reuters or Associate Press). These news websites are maintained by large, professional news companies and are updated several times a day.

4.3.1 Data Description

Most news websites provide RSS feeds that list all new articles in chronological order. Originally, they were created to keep their readers updated, but they can also be used by an automated process to download new content from those websites.

The data for this case study comes from the test and evaluation stages of the process at Siemens Corporate Technology. In this context we were able to use real world test cases, which are unfortunately of proprietary nature by the time of writing. It has been our goal to perform a real world case study to test the feasibility and consequent usability in a corporate environment.

In the course of this case study news articles were downloaded from popular and well-frequented commercial sites providing business related news. Some of the websites offer more than one category concerned with business specific topics. For these, all categories were chosen. Table 4.7 provides a concise overview of all sources that were used.

This section presents the results of the clustering of news articles and is based on a data set of over 2,200 articles which were published from February, 2nd until March, 21st 2006. All of the news websites used for this case study are operated commercially. Term

Provider	Selected Category
ABC News	Business Headlines
BBC News	E-Commerce: World Headlines Economy: World edition Business: UK Edition Business: World Edition
CBC News	Business News
CNN Money	News/Economy
MSNBC	Business
Telegraph Co.	Business Small Business Markets Personal Finance
Yahoo! News	Business

Table 4.7: Sources of news articles that were used for this case study.

selection reduced the number of terms from 5,873 to 1,959. The total number of words was (after filtering) 38,278. The distribution of terms with the heaviest occurrence can be seen in 4.9(b). For the visualization in 4.10 a subset consisting of 303 articles was used.

4.3.2 Time Frame Selection

The selection of a two-tier hierarchy was based upon the activity of postings and news articles. A lower tier foreground model uses a 7 day time-frame, and the second higher tier uses 14 days. This underscores the intention of tracking longer term trends with the second tier, while using the first tier to find localized topic instances. Figure 4.9(a) shows the observed activity of covered data sources in the time between February 04 and March 21 of last year. The peaks lie approximately 7 days apart, including a two-day hiatus over the weekends. This basic rhythm is used to determine the topic life-cycle lengths, i.e. weekly and bi-weekly topics. The analysis can be readily expanded to include larger time-frames, this is in fact an area of further interest to us.

Nr	Source	Headline
1	Telegraph	Followers of New Star can find more sparklers above £1bn
	ABC	Oil Prices Rise to Just Over $61 a Barrel
	ABC	Abqaiq Facility at Heart of Saudi Industry
	ABC	Monsanto Settles Lawsuit Over Hormone
	ABC	Oil, Gasoline Futures Fall Sharply
	MSNBC	Ferrari driving school coming to North America
	BBC	Saudis ‘foil oil facility attack’
	Yahoo!	Oil climbs 4 percent on Saudi attack
	Yahoo!	Al Qaeda vows more attacks after Saudi oil raid
2	MSNBC	Venezuela to reduce some U.S. airline flights
	BBC	Venezuela cuts US airline flights
	Yahoo!	Venezuela delays US flight curbs for more talks
	BBC	Venezuela cuts US airline flights
3	ABC	Enron Witness Says He Plundered Reserves
	CBS	Enron accountant says he fudged number to please CEO
	MSNBC	Ex-Enron accountant says he cooked books

Table 4.8: Sample result clusters

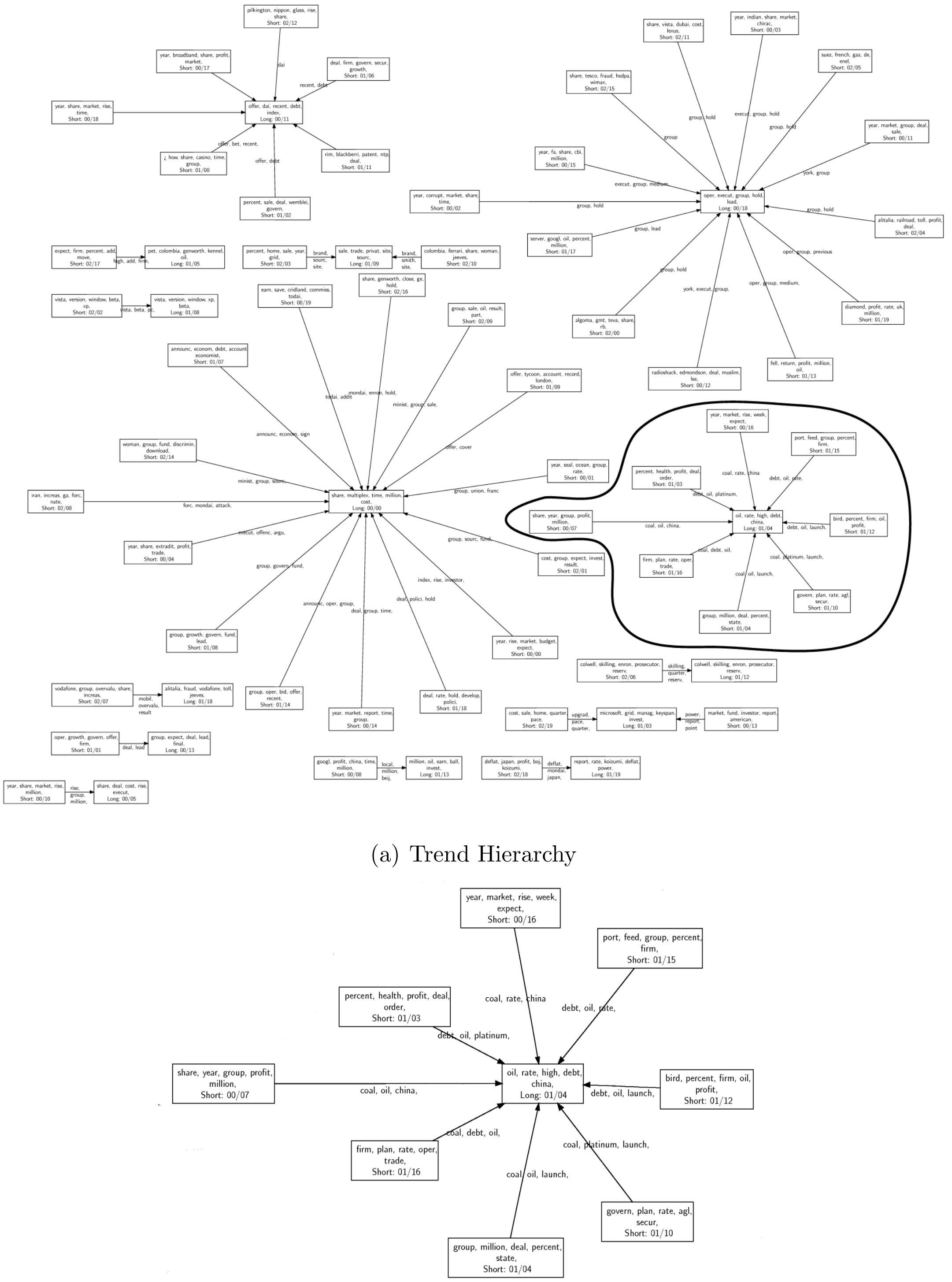
4.3.3 Results

The resulting clustering on the long-term trend level (i.e. FG_l) contained 40 clusters. Three samples of high quality clusters can be seen in table 4.8, but almost 60% of the clusters had to be considered trash clusters. Discounting the noise from the result set leaves a handful of clusters with well defined contents.

Figure 4.10(a) shows a visualization of the topic hierarchy extracted from the data. Every box represents a cluster. The cluster at the center of a circular structure represents a long cluster representing a general trend, those around it short clusters of specific topics.

Each cluster is marked with the five most typical words in comparison to the background model. Edges are annotated with the words that both clusters share. The spelling of words is caused by the stemming process during preprocessing. The first number of a cluster denotes an internal index of its time period, whereas the second number stands for the index of the cluster for this particular interval.

One hierarchy is highlighted in 4.10(b). It shows several short clusters that are connected to large cluster 01/04. The short clusters originate from several time periods and discuss a common topic as can be seen from the typical words of all clusters involved.


pilkington, nippon, glass, rise, share, Short: 02/12
year, broadband, share, profit, market, Short: 00/17
deal, firm, govern, secur, growth, Short: 01/06
dai
recent, debt
year, share, market, rise, time, Short: 00/18
offer, dai, recent, debt, index, Long: 00/11
offer, bet, recent,
offer, debt
¿ how, share, casino, time, group, Short: 01/00
rim, blackberri, patent, ntp, deal, Short: 01/11
percent, sale, deal, wemblei, govern, Short: 01/02
share, vista, dubai, cost, lexus, Short: 02/11
year, indian, share, market, chirac, Short: 00/03
suez, french, gaz, de, enel, Short: 02/05
share, tesco, fraud, hsdpa, wimax, Short: 02/15
year, market, group, deal, sale, Short: 00/11
group, hold
execut, group, hold
group, hold
group
york, group
year, fa, share, cbi, million, Short: 00/15
execut, group, medium
year, corrupt, market, share, time, Short: 00/02
group, hold
oper, execut, group, hold, lead, Long: 00/18
group, hold
alitalia, railroad, toll, profit, deal, Short: 02/04
group, lead
server, googl, oil, percent, million, Short: 01/17
oper, group, previous
group, hold
york, execut, group,
oper, group, medium,
algoma, gmt, teva, share, rb, Short: 02/00
diamond, profit, rate, uk, million, Short: 01/19
radioshack, edmondson, deal, muslim, lse, Short: 00/12
fell, return, profit, million, oil, Short: 01/13
expect, firm, percent, add, move, Short: 02/17
high, add, firm,
pet, colombia, genworth, kennel, oil, Long: 01/05
percent, home, sale, year, grid, Short: 02/03
brand, sourc, site,
sale, trade, privat, site, sourc, Long: 01/09
brand, smith, site,
colombia, ferrari, share, woman, jeeves, Short: 02/10
vista, version, window, beta, xp, Short: 02/02
vista, beta, pc,
vista, version, window, xp, beta, Long: 01/08
earn, save, cridland, commiss, todai, Short: 00/19
share, genworth, close, ge, hold, Short: 02/16
group, sale, oil, result, part, Short: 02/09
announc, econom, debt, account, economist, Short: 01/07
offer, tycoon, account, record, london, Short: 01/09
todai, addit
mondai, enron, hold,
minist, group, sale,
announc, econom, sign
offer, cover
woman, group, fund, discrimin, download, Short: 02/14
year, seal, ocean, group, rate, Short: 00/01
minist, group, sourc,
group, union, franc
iran, increas, ga, forc, nate, Short: 02/08
forc, mondai, attack,
share, multiplex, time, million, cost, Long: 00/00
execut, offenc, argu,
year, share, extradit, profit, trade, Short: 00/04
group, sourc, fund,
cost, group, expect, invest, result, Short: 02/01
group, govern, fund,
index, rise, investor,
announc, oper, group,
deal, polici, hold
deal, group, time,
group, growth, govern, fund, lead, Short: 01/08
year, rise, market, budget, expect, Short: 00/00
group, oper, bid, offer, recent, Short: 01/14
year, market, report, time, group, Short: 00/14
deal, rate, hold, develop, polici, Short: 01/18
year, market, rise, week, expect, Short: 00/16
port, feed, group, percent, firm, Short: 01/15
percent, health, profit, deal, order, Short: 01/03
coal, rate, china
debt, oil, rate,
debt, oil, platinum,
share, year, group, profit, million, Short: 00/07
coal, oil, china,
oil, rate, high, debt, china, Long: 01/04
debt, oil, launch,
bird, percent, firm, oil, profit, Short: 01/12
coal, debt, oil,
firm, plan, rate, oper, trade, Short: 01/16
coal, platinum, launch,
coal, oil, launch,
govern, plan, rate, agl, secur, Short: 01/10
group, million, deal, percent, state, Short: 01/04
colwell, skilling, enron, prosecutor, reserv, Short: 02/06
skilling, quarter, reserv,
colwell, skilling, enron, prosecutor, reserv, Long: 01/12
vodafone, group, overvalu, share, increas, Short: 02/07
mobil, overvalu, result
alitalia, fraud, vodafone, toll, jeeves, Long: 01/18
cost, sale, home, quarter, pace, Short: 02/19
upgrad, pace, quarter,
microsoft, grid, manag, keyspan, invest, Long: 01/03
power, report, point
market, fund, investor, report, american, Short: 00/13
oper, growth, govern, offer, firm, Short: 01/01
deal, lead
group, expect, deal, lead, final, Long: 00/13
googl, profit, china, time, million, Short: 00/08
local, million, beij,
million, oil, earn, ball, invest, Long: 01/13
deflat, japan, profit, boj, koizumi, Short: 02/18
deflat, mondai, japan,
report, rate, koizumi, deflat, power, Long: 01/19
year, share, market, rise, million, Short: 00/10
rise, group, million,
share, deal, cost, rise, execut, Long: 00/05
year, market, rise, week, expect, Short: 00/16
port, feed, group, percent, firm, Short: 01/15
coal, rate, china
debt, oil, rate,
percent, health, profit, deal, order, Short: 01/03
debt, oil, platinum,
share, year, group, profit, million, Short: 00/07
coal, oil, china,
oil, rate, high, debt, china, Long: 01/04
bird, percent, firm, oil, profit, Short: 01/12
debt, oil, launch,
coal, debt, oil,
firm, plan, rate, oper, trade, Short: 01/16
coal, platinum, launch,
coal, oil, launch,
govern, plan, rate, agl, secur, Short: 01/10
group, million, deal, percent, state, Short: 01/04

(a) Trend Hierarchy

(b) Sample Result

Figure 4.10: Sample results of the hierarchical tracking approach. Result clusters from a long time period are shown in the circular structures.

Every box represents a cluster. The cluster at the center of a circular structure represents a long cluster, those around it short clusters. Each cluster is marked with the five most typical words in comparison to the background model. Edges are annotated with the words that both clusters share. The spelling of words is caused by the stemming process during preprocessing. The first number of a cluster denotes an internal index of its time period, whereas the second number stands for the index of the cluster for this particular interval.

Figure 4.10 shows that it is possible to find relations between clusters of shorter time periods and more general topics present for a much larger time scale.

4.4 Outlook and Conclusion

During this work a number of issues have at time been considered but not followed up on. Particularly one idea pertaining to graph clustering had been considered, but when it did not present the expected results in conjunction with the other analyses it was considered to be continuing work.

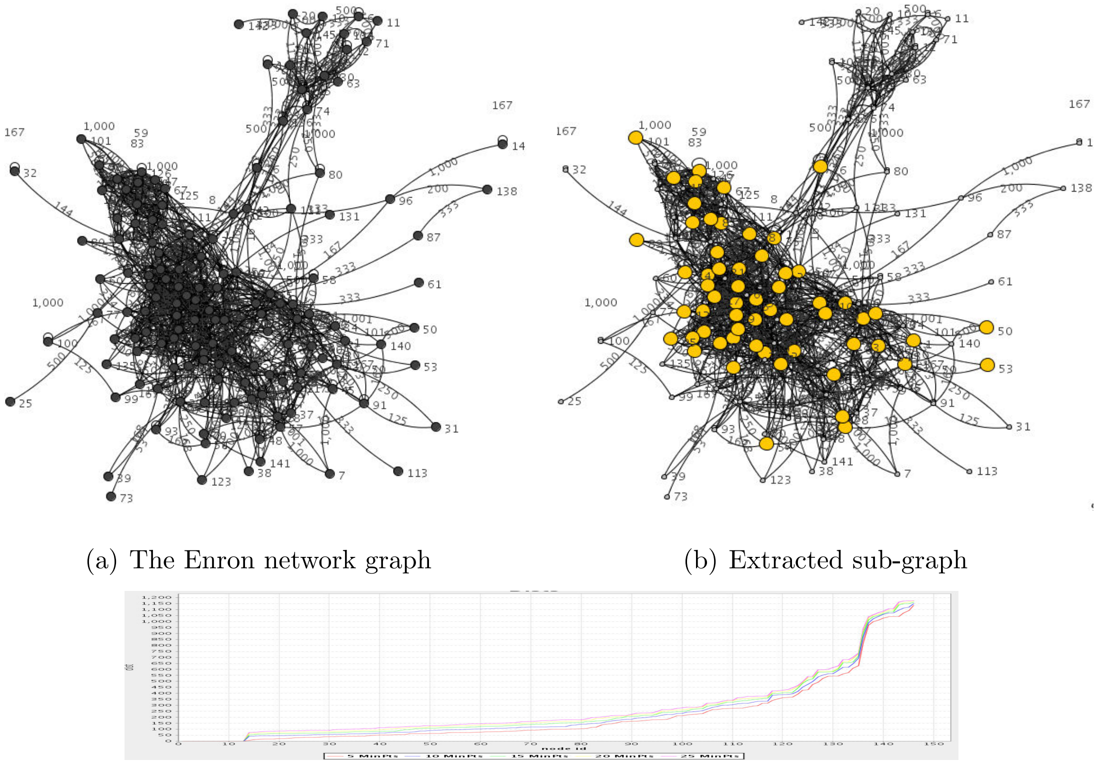
4.4.1 Partitioning Graph Networks

Network analysis, particularly social network analysis relies on the correct identification of sub-networks as to present a semantically coherent subset to analyze. Generally the data gathering step is used to filter the information in a preprocessing step or to limit the scope of considered information in such a way as to produce semantically sensible graphs. When one is faced with massive graphs, on the other hand, it becomes necessary to extract such sub-networks from the basic graph data. It is our proposed approach to use density of a graph as defined by link distance as a basis for the extraction of dense sub-graphs.

We have applied the use of density based clustering with noise detection (both the DBSCAN [42] and OPTICS [9] algorithms) to graphs. While the results were interesting, they were not conclusive in the context of social network analysis, and as such will be pursued in the future.

Enron

The Enron network consists of solely 146 actors, and as such is not very large. While the clustering did perform adequately, the only cluster found is the one large central lump centered around the few central actors (Figure 4.11(a) and 4.11(b)). While we regard this as interesting, it did not avail a new insight into the partitioning graphs to support social network analysis.



(a) The Enron network graph

(b) Extracted sub-graph

(c) Nearest Neighborhood Graph

Figure 4.11: Density based clustering of graphs

4.4.2 Challenges

The data sets in question were quite large and complex, but every set of real world data presents its own caveats. We ran into numerous instances which necessitated particular tuning of our approach. This indicates a great deal of exploration of the quirks must be made if a large-scale application should be developed from these approaches.

Text mining, specifically the use of keywords as a basis for a vector space representation of documents is still more an art than a science. Particularly the prevalence of informal spelling in short textual messages presents a significant impediment to the use of automated keyword extraction techniques. While the use of dictionaries and possible words within a short Levenshtein distance can correct numerous spelling errors, the use of idioms and colloquial terms increases the difficulty of parsing and analyzing text without human supervision.

The Enron data set also presented a pervasive problem faced in real world data, namely that of incompleteness. As was noted in the section 4.1 the data present is most likely a collection in-boxes found on email servers at the time of the investigation. This data most likely represents only a part of the email network, which must be captured at the source if the demand for completeness is to be satisfied.

4.4.3 Conclusion

The approaches put forward in Section 3.3 have been put to the test using real world data. The results indicate the utility of dissecting large communications corpora to increase the accuracy of social network analysis of these corpora.

We have taken the inherent assumption of **social network analysis**, namely the disregard of content when judging **centrality** and **prestige** of actors in a social network, into consideration when analyzing massive graphs. The idea of calculating the centrality of a large, **semantically not homogeneous network**, dilutes the expressiveness of social network analysis to the degree of generality. This has lead us to take content into account when performing social network analysis. By focusing on the topics contained in communication networks, we can discover and extract **topic based sub-networks** from massive graphs.

Such a topic based network, when treated to social network analysis, now describes centrality and prestige in a much more precise setting than the entire graph. In our case we find that a broad content cluster concerned with office humour tends to be spread not by the management, but rather by actors found lower in the corporate hierarchy. This is by no means a clear-cut rule, as even a CEO may find a joke worth sending on.

By joining the areas of topic analysis and social network analysis, we can now segment a communication corpus based on text messages into topics, each of which can be analyzed separately to yield **topic based centrality** values. Whereas the result is not as pronounced as in the Enron data set, we have found relevant cohesive sub-networks in both cases. The resulting increased expressiveness of topic based centrality has consequently yielded a better understanding of the communications taking place within the complete communication networks.

Discovery and tracking of topics in text corpora has previously focused on dissecting the time period into sliding windows of arbitrary size, and subsequently guessing at the correct window size and overlap which critically impact the recognition of topics in their life-cycle. Thus the topics recognized are greatly dependant upon the selection of parameters defining window size and extent. In order to compare topics across differing time frames with respect to the overall text corpora, we are able to use **representative keyword vectors** to sum up topics by contrasting the topics occurring within a foreground model with the general background model.

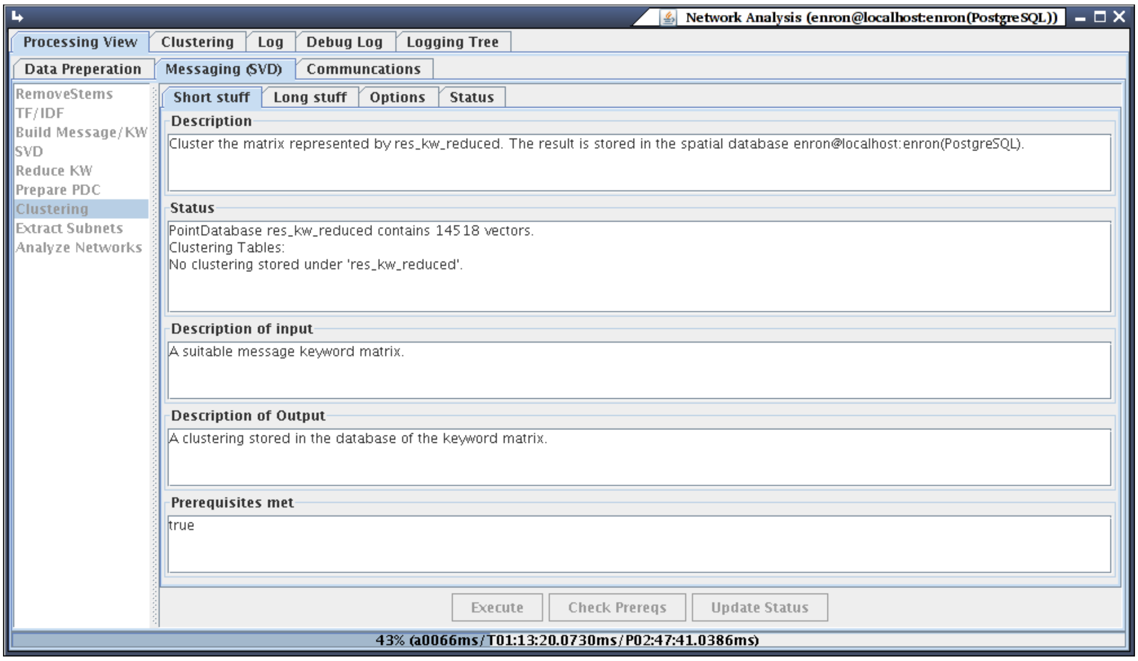
By introducing a multi-tier temporal hierarchy, we have been able to follow topic trends as the evolution of long term topic clusters, expressed by the emergence and disappearance of short-term topics linked to the long term trend. Within these evolving topics the used of focused social network analysis can generated an insight into opinion leaders and topic related heavy hitters.

Furthermore the use of Eigenvector based social network analysis can also lead to the discovery of relevant topics and central actors. These dynamics would have applications for example in detecting shifts of interest or the rise of new 'buzz words' in a given community. This could be revealed in at least two ways: first by the analysis of the authors-use-of-words when the relevant players have already been identified. Thus the 'spin-doctors' and the 'pet-topics' would become visible.This could be relevant for example for the development of innovative products and services if such an analysis could be performed to 'keep an eye' on topics in R&D communities.

5 Implementations

Contents

This thesis has touched upon numerous computational aspects throughout the past few chapters. This chapter will now go into greater detail where the algorithms and computational framework underlying this thesis are concerned. The calculations performed during the course of the work have been mentioned where appropriate, and this chapter will detail the effort and complexities faced over time. The numerous computational steps necessary to perform the analyses as delineated in this thesis motivated a sizable effort in terms of implementation and maintenance. This chapter gives an overview of the challenges faced during the execution of the analyses as well as the overarching structure of the process used.



Description

Cluster the matrix represented by res_kw_reduced. The result is stored in the spatial database enron@localhost:enron(PostgreSQL).

Status

```
PointDatabase res_kw_reduced contains 14518 vectors.
Clustering Tables:
No clustering stored under 'res_kw_reduced'.
```

Description of input

A suitable message keyword matrix.

Description of Output

A clustering stored in the database of the keyword matrix.

Prerequisites met

true

43% (a0066ms/T01:13:20.0730ms/P02:47:41.0386ms)

Figure 5.1: Application Screenshot

5.1 Introduction

In order to create a more seamless working environment a number of analyses were combined to form a common analysis tool incorporating most aspects of the computational work. This application relies heavily upon a computational framework constructed during this thesis as well as external libraries. The computational platform consists of a combined interface window shown in Figure 5.1. This screenshot shows the initial data preparation stage for the text mining of communications corpora.

During the course of this work a a portion of the computational analysis was performed by analytical software of varied origins. While a number of external analysis projects are established software suites, we also relied on analytical libraries and implemented code to surmount the different challenges presented by the approaches put forward in this work. We will differentiate between three categories of analytics:

- Established software suites designed to perform specific calculations in an efficient and effective manner. Examples include Mathematica [129] and the statistical R [110] packages.
- Software libraries offering analytical support for specific problems such as data visualization (JFreechart[27], Java Universal Network/Graph Framework [93]) and linear algebra (COLT[23]).
- Implementations of algorithms in the areas of social network analysis and text mining.

It is the last category of analytics we will now talk about, as well as the framework tying everything together. Beyond the implementations affected during the course of this thesis a number of computational aspects relevant to social network analysis and text mining (or more specifically topic discovery) will form an overview of the practical aspects confronting social network analysis today.

At the center of any analyses performed during this thesis lies a collected set of algorithms and frameworks to provide flexible and powerful analytical tools. From the treatment of text data to extract keyword vectors to the social network analysis to be performed on communication graphs, a varied set of tools were needed to bring social network analysis together with text mining and clustering.

As a result a flexible basis for performing the various steps has grown from simple tasks (e.g. performing TF/IDF) to the integration of steps into a set of measures to scan, stem and process a set of text messages into a normalized keyword vector space. Figure 5.2 shows a structural diagram of the software designed.

It was clear from the beginning that any analysis will have to be run repeatedly with varying parameters. For this reason we concentrated on a structure which facilitated the creation of tightly defined independent steps designed to execute various aspects of the larger goal of partitioning massive graphs. The application was thus designed to act as an execution and test-bed for the various **processing steps**. To add some structure the bag of processing steps a number of such processing steps can be combined into a **processing group**. Each processing group receives its own viewing pane. The contract for a processing step provides for feedback as well as input for the individual step. Each step must also work independently of any other processing step, i.e. reading the input data from the database and writing a defined result set back into the database.

There are two processing groups to note, one to perform data import and preparation, and a second to perform the actual analysis geared towards social network analysis of massive graphs. The primary reason for this divide was the idea of applying the analysis (combined into **DataAnalysis** in Figure 5.2) to multiple data sets, each of which demanding a slightly different method to import and clean raw data as well as slightly diverging methods used for data preparation.

Processing Group: Data Preperation

This group is concerned with reading flat data, such as a set of emails as is the case with the Enron data set, or other unstructured data into a database. This ensures at least a uniform access and storage of data. By performing preparatory functions such as the extraction of a keyword vector space in the case of text or the recognition of named entities, the groundwork for the two aspects (structure and content of communications corpora) this thesis sets out to explore is laid.

Figure 5.2 shows the text analysis preparation in somewhat greater detail, as the number of filtering steps and stemming steps are all conceived as implementations of a generic **ParsingStep**. One common hurdle to text analysis is the occurrence of multiple languages, but as can be seen we only needed two dictionaries, namely **StopWordsEN**

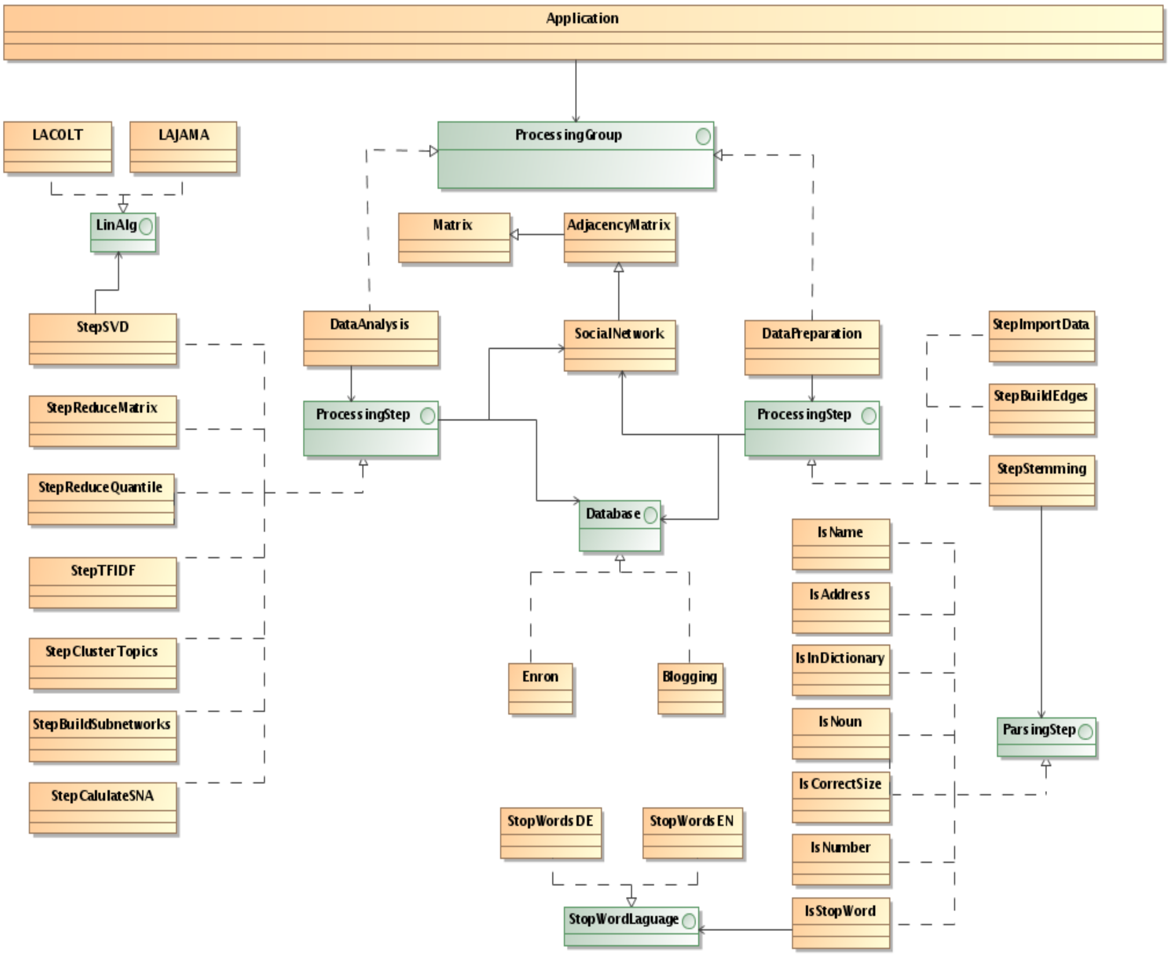


Figure 5.2: Structure Flowchart

and **StopWordsDE** for the individual treatment of stop-words in text. Stemming was performed independently for each language. The final product is a cleaned data set represented by a set of keyword vectors and an associated communications graph (somewhat generically named a **SocialNetwork**).

Processing Group: Data Analysis

After data preparation generates a clean initial data set, this processing group is designed to run the various aspects of the massive graph dissection process. By taking the preprocessed data from the database the social network analysis of massive graphs is pursued: the keyword vectors are normalized, pruned and reduced (through the use of steps **StepTFIDF**, **StepReduceQuantile**, **StepSVD** and **StepReduceMatrix** respectively). As can be seen at the bottom of Figure 5.2, the use of two linear alge-

Java Package	Description
net.viermetz.clustering	Package containing clustering related classes, such as the clustering plug-in framework, and of course the individual classes implementing the various clustering algorithms.
net.viermetz.data	Data definitions such as vectors, etc.
net.viermetz.graph	Graph related data structures such as edges and nodes, as well as graph related analytical algorithms.
net.viermetz.index	Index structures for data, notably the SR-Tree for clustering support.
net.viermetz.logging	Logging facilities, visual representation of logs and logging trees and custom sinks.
net.viermetz.math	Basic classes describing vector computation, matrix operations and the like.
net.viermetz.rendering	Visualization package for vector spaces.
net.viermetz.text	Text analysis classes such as stemming, TF/IDF and related algorithms.

Table 5.1: Packages providing peripheral computational support

bra packages has been implemented (**LACOLT** slots the COLT [23] packages into the application, and **LAJAMA** uses the JAMA package [60].).

The resulting reduced keyword space is then clustered for topics, in this case using a set of clustering algorithms not detailed in the figure as they fall into the Framework discussed in the following section. From these results the sub-networks can then be extracted and targeted, or content related, social network analysis be performed.

5.2 Structure and Framework

A great deal of the implementation work was performed in the initial phase of this thesis, therefore the package structure reflects the **net.viermetz.*** nomenclature. While both the application, as just discussed, as well as the framework of supporting packages it is built upon, share the same package tree. A concise overview of just this framework will now be given.

5.2.1 Framework

A number of constructs have been used which do not directly influence this dissertation in a specific context, but which provide a necessary background of capabilities. These range from simple data structures to graphical support classes. All of these classes have been grouped into a background framework designed to offer maximum flexibility with a minimum of programming effort. Table 5.1 lists a few of the more relevant packages supporting the analyses in this thesis.

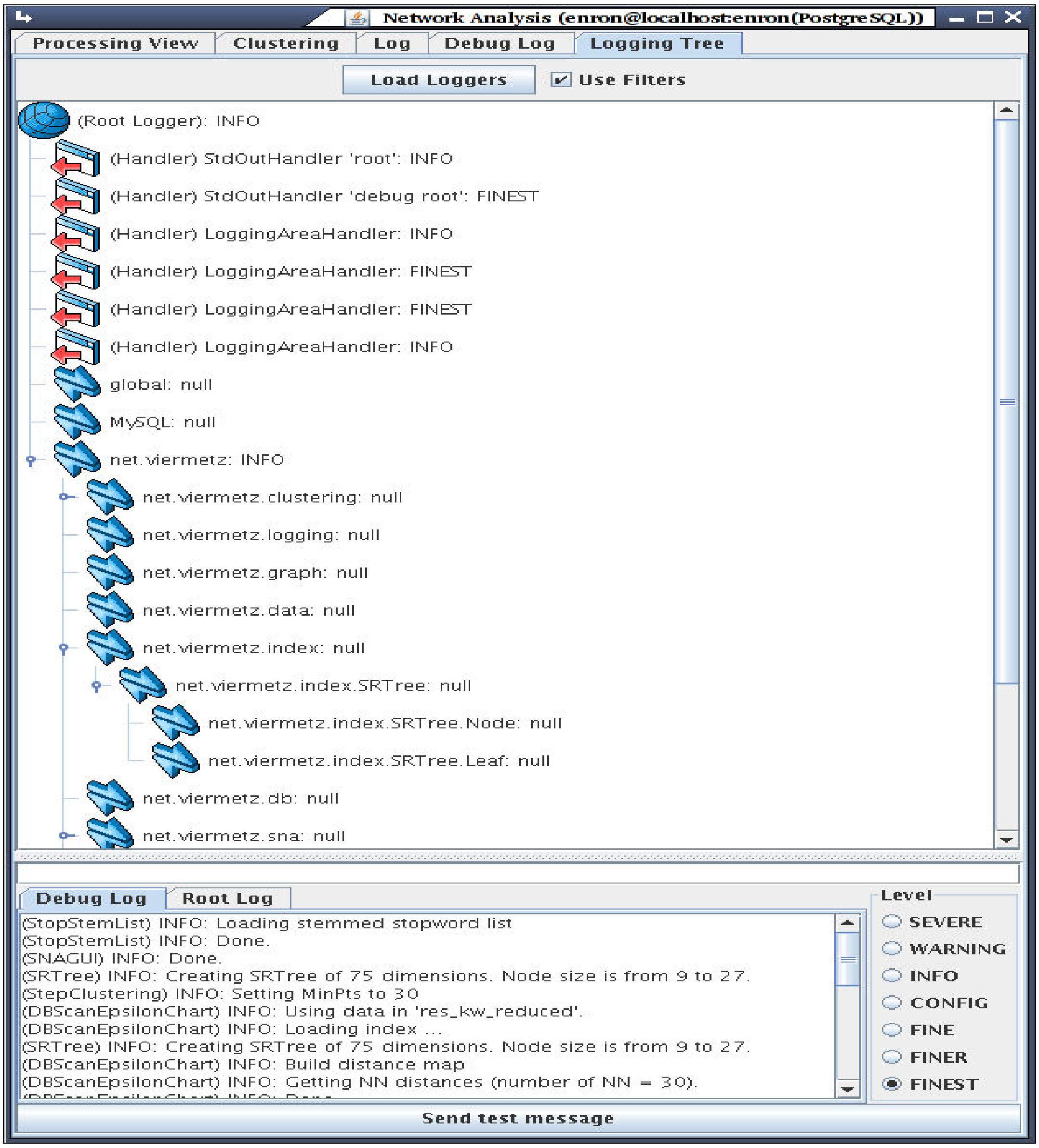
A goal was to implement when necessary, and to reuse libraries when possible. As the overall integration became a focus of the programming perspective, only low level libraries providing for instance linear algebra support has been integrated, as the maturity of third party packages was seldom developed to a satisfying extent. For instance while numerous linear algebra packages exist, often showing vastly differing states of maturity, existing solutions show little of the flexibility needed in a more complex environment, e.g. a clustering framework fitting over spatial as well as graph based data. For this reason a number of custom fit solutions were pursued, which will be discussed on the following pages.

5.2.2 Application

The combined analytics package is served by a GUI designed to allow access to all stages during communications corpus segmentation. Shown in Figure 5.3, the general layout divides the functionality into three basic categories, namely Corpora Processing (the *processing* tab), a Clustering and Topic Discovery (denoted by the *clustering* tab) and the three logging tabs. The logging uses the *java.util.logging* package, which is displayed visually in Figure 5.3. Here the various logging levels and sinks can be adjusted interactively, for instance allowing a more detailed output of a clustering run by increasing the verbosity of the *net.viermetz.clustering* sink.

The next few pages will detail the inner workings of the Corpora Processing as well as the Clustering and Topic Discovery processing stages. Social network analysis is performed within the Corpora Processing stage, but is mentioned here on its own.

Preliminary to any analysis the basic data has to be parsed and brought into the expected format and database repository. The importing and parsing is shown in Figure 5.4, and is only responsible for the proper and complete surveying and storing of the given communications data. For instance in the case of the Enron network this entailed reading flat text files and constructing a name register and communications network from the to and from address fields of the individual emails.



☑ Use Filters

```
(Root Logger): INFO
    (Handler) StdOutHandler 'root': INFO
    (Handler) StdOutHandler 'debug root': FINEST
    (Handler) LoggingAreaHandler: INFO
    (Handler) LoggingAreaHandler: FINEST
    (Handler) LoggingAreaHandler: FINEST
    (Handler) LoggingAreaHandler: INFO
    global: null
    MySQL: null
    net.viermetz: INFO
        net.viermetz.clustering: null
        net.viermetz.logging: null
        net.viermetz.graph: null
        net.viermetz.data: null
        net.viermetz.index: null
            net.viermetz.index.SRTree: null
                net.viermetz.index.SRTree.Node: null
                net.viermetz.index.SRTree.Leaf: null
        net.viermetz.db: null
        net.viermetz.sna: null
```

```
(StopStemList) INFO: Loading stemmed stopword list
(StopStemList) INFO: Done.
(SNAGUI) INFO: Done.
(SRTree) INFO: Creating SRTree of 75 dimensions. Node size is from 9 to 27.
(StepClustering) INFO: Setting MinPts to 30
(DBScanEpsilonChart) INFO: Using data in 'res_kw_reduced'.
(DBScanEpsilonChart) INFO: Loading index ...
(SRTree) INFO: Creating SRTree of 75 dimensions. Node size is from 9 to 27.
(DBScanEpsilonChart) INFO: Build distance map
(DBScanEpsilonChart) INFO: Getting NN distances (number of NN = 30).
```

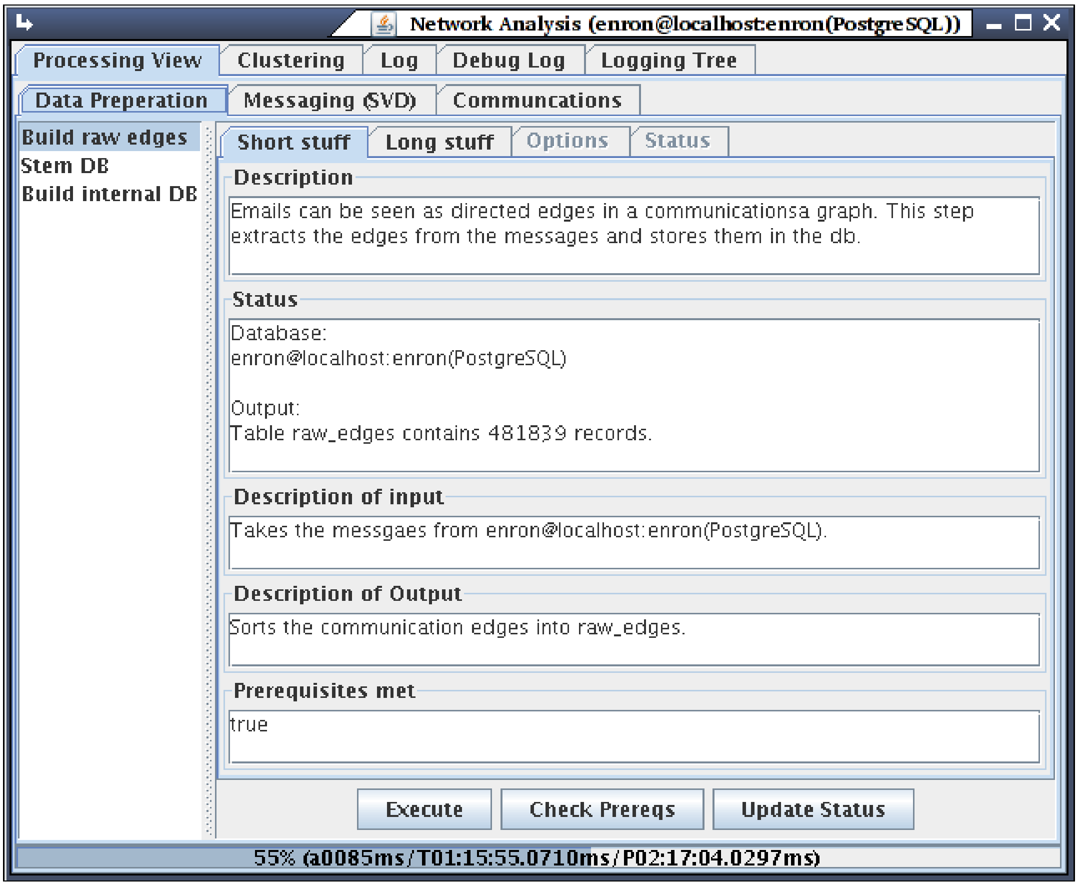
Level
○ SEVERE
○ WARNING
○ INFO
○ CONFIG
○ FINE
○ FINER
● FINEST

Figure 5.3: The logging facility

Text Mining

The basic steps to prepare the data into keyword vector format are all performed in the Corpora Processing Tab. Figure 5.5(a) shows a sample message to be treated to the process. The process can be stepped through incrementally to show the sequential effects of the filtering and stemming to be done.

After the data has been imported and cleaned of superfluous information (such as message wrappers and attached files), the actual processing can start. As can be seen in Figure 5.5(b) preliminary text filtering removes any extraneous punctuation and gram-



Description

Emails can be seen as directed edges in a communicationsa graph. This step extracts the edges from the messages and stores them in the db.

Status

Database:
enron@localhost:enron(PostgreSQL)

Output:
Table raw_edges contains 481839 records.

Description of input

Takes the messgaes from enron@localhost:enron(PostgreSQL).

Description of Output

Sorts the communication edges into raw_edges.

Prerequisites met

true

55% (a0085ms/T01:15:55.0710ms/P02:17:04.0297ms)

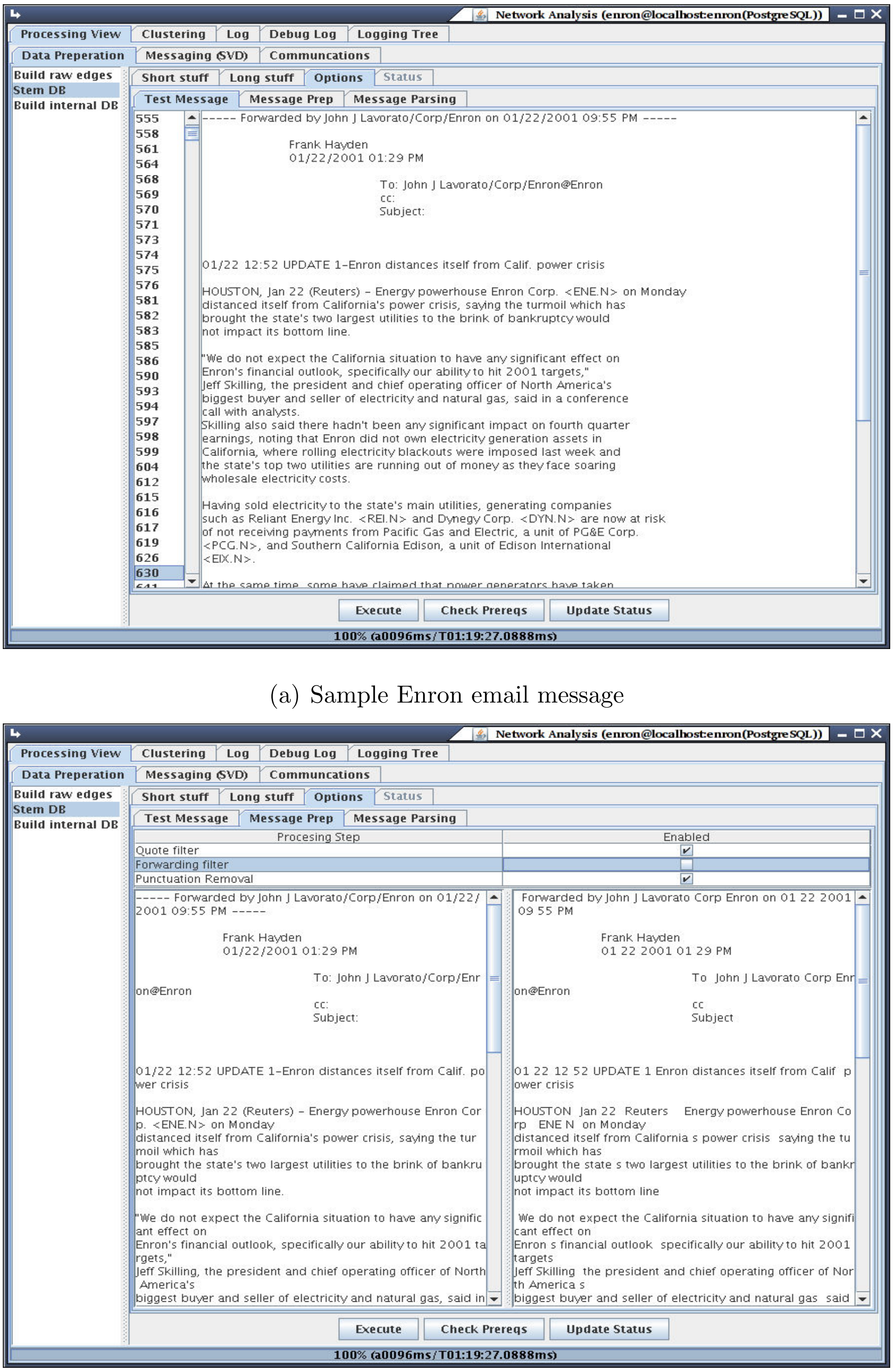
Figure 5.4: Parse Communications Data

mar.

Finally the keywords in the database are compared to blacklists and white-lists (see Figure 5.6) before being treated to stemming. Words not being rejected are shown in following columns, a red mark depicting any word which has been removed or discarded during the process. Any stem arriving at the end of the processing pipe is displayed in the far right column with a green background. These stems form the basis of the keyword vector space used for topic discovery and content analysis.

An important step not depicted here is the normalization of keyword vectors over frequency and use by performing TF/IDF on all stems in the database. The normalized vector data is also treated to pruning of too frequent or infrequent terms, in our case a long tail of many infrequent terms meant that a reduction of terms contributing less than a tenth of a percent to the most common contributor includes a sizable portion of terms, almost 15% to 29% varying from data set to data set.

The resulting vector space, pruned and normalized, is now considered to be the basis for further treatment to clustering of topics and also social network analysis.



```
555
558
561
564
568
569
570
571
573
574
575
576
581
582
583
585
586
590
593
594
597
598
599
604
612
615
616
617
619
626
630
```

```
----- Forwarded by John J Lavorato/Corp/Enron on 01/22/2001 09:55 PM -----

                    Frank Hayden
                    01/22/2001 01:29 PM

                                        To: John J Lavorato/Corp/Enron@Enron
                                        cc:
                                        Subject:



01/22 12:52 UPDATE 1-Enron distances itself from Calif. power crisis

HOUSTON, Jan 22 (Reuters) - Energy powerhouse Enron Corp. <ENE.N> on Monday
distanced itself from California's power crisis, saying the turmoil which has
brought the state's two largest utilities to the brink of bankruptcy would
not impact its bottom line.

"We do not expect the California situation to have any significant effect on
Enron's financial outlook, specifically our ability to hit 2001 targets,"
Jeff Skilling, the president and chief operating officer of North America's
biggest buyer and seller of electricity and natural gas, said in a conference
call with analysts.
Skilling also said there hadn't been any significant impact on fourth quarter
earnings, noting that Enron did not own electricity generation assets in
California, where rolling electricity blackouts were imposed last week and
the state's top two utilities are running out of money as they face soaring
wholesale electricity costs.

Having sold electricity to the state's main utilities, generating companies
such as Reliant Energy Inc. <REI.N> and Dynegy Corp. <DYN.N> are now at risk
of not receiving payments from Pacific Gas and Electric, a unit of PG&E Corp.
<PCG.N>, and Southern California Edison, a unit of Edison International
<EIX.N>.

At the same time, some have claimed that power generators have taken
```

(a) Sample Enron email message

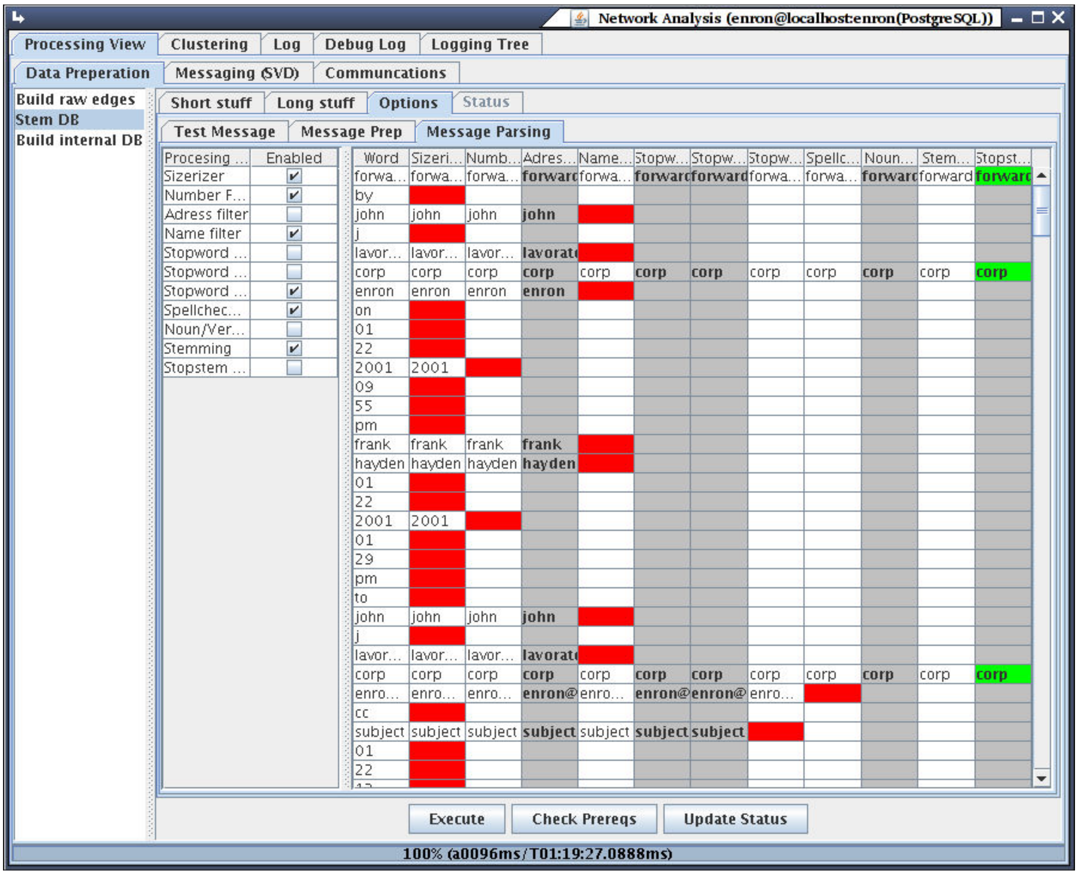
Procesing Step	Enabled
Quote filter	☑
Forwarding filter	☐
Punctuation Removal	☑

```
----- Forwarded by John J Lavorato/Corp/Enron on 01/22/
2001 09:55 PM -----

             Frank Hayden
             01/22/2001 01:29 PM

                         To: John J Lavorato/Corp/Enr
on@Enron
                         cc:
                         Subject:



01/22 12:52 UPDATE 1-Enron distances itself from Calif. po
wer crisis

HOUSTON, Jan 22 (Reuters) - Energy powerhouse Enron Cor
p. <ENE.N> on Monday
distanced itself from California's power crisis, saying the tur
moil which has
brought the state's two largest utilities to the brink of bankru
ptcy would
not impact its bottom line.

"We do not expect the California situation to have any signific
ant effect on
Enron's financial outlook, specifically our ability to hit 2001 ta
rgets,"
Jeff Skilling, the president and chief operating officer of North
 America's
biggest buyer and seller of electricity and natural gas, said in
```

```
 Forwarded by John J Lavorato Corp Enron on 01 22 2001
09 55 PM

             Frank Hayden
             01 22 2001 01 29 PM

                         To  John J Lavorato Corp Enr
on@Enron
                         cc
                         Subject



01 22 12 52 UPDATE 1 Enron distances itself from Calif  p
ower crisis

HOUSTON  Jan 22  Reuters    Energy powerhouse Enron Co
rp   ENE N  on Monday
distanced itself from California s power crisis  saying the tu
rmoil which has
brought the state s two largest utilities to the brink of bankr
uptcy would
not impact its bottom line

 We do not expect the California situation to have any signifi
cant effect on
Enron s financial outlook  specifically our ability to hit 2001
targets
Jeff Skilling  the president and chief operating officer of Nor
th America s
biggest buyer and seller of electricity and natural gas  said
```

(b) Test Message Selection

Figure 5.5: Text message parsing



Procesing …	Enabled
Sizerizer	☑
Number F…	☑
Adress filter	□
Name filter	☑
Stopword …	□
Stopword …	□
Stopword …	☑
Spellchec…	☑
Noun/Ver…	□
Stemming	☑
Stopstem …	□

Word	Sizeri…	Numb…	Adres…	Name…	Stopw…	Stopw…	Stopw…	Spellc…	Noun…	Stem…	Stopst…
forwa…	forwa…	forwa…	forwarc	forwa…	forwarc	forward	forwa…	forwa…	forwarc	forward	forwarc
by	🟥										
john	john	john	john	🟥							
j	🟥										
lavor…	lavor…	lavor…	lavorat	🟥							
corp	corp	corp	corp	corp	corp	corp	corp	corp	corp	corp	corp
enron	enron	enron	enron	🟥							
on	🟥										
01	🟥										
22	🟥										
2001	2001	🟥									
09	🟥										
55	🟥										
pm	🟥										
frank	frank	frank	frank	🟥							
hayden	hayden	hayden	hayden	🟥							
01	🟥										
22	🟥										
2001	2001	🟥									
01	🟥										
29	🟥										
pm	🟥										
to	🟥										
john	john	john	john	🟥							
j	🟥										
lavor…	lavor…	lavor…	lavorat	🟥							
corp	corp	corp	corp	corp	corp	corp	corp	corp	corp	corp	corp
enro…	enro…	enro…	enron@	enro…	enron@	enron@	enro…	🟥			
cc	🟥										
subject	subject	subject	subject	subject	subject	subject	🟥				
01	🟥										
22	🟥										

100% (a0096ms/T01:19:27.0888ms)

Figure 5.6: Stemming

5.2.3 Clustering

The goal of clustering spatial data as well as graph data, coupled with the requirement to generate synthetic test data has led us to implement a small number of clustering approaches within our framework. We decided to implement the SR-Tree[74] as an index structure, and did not need to fall back on any other data structure. This index was chosen to facilitate the greatest and simplest access to spatial data sets ranging from a few hundred to roughly half a million data vectors spanning anything up to several thousand dimensions.

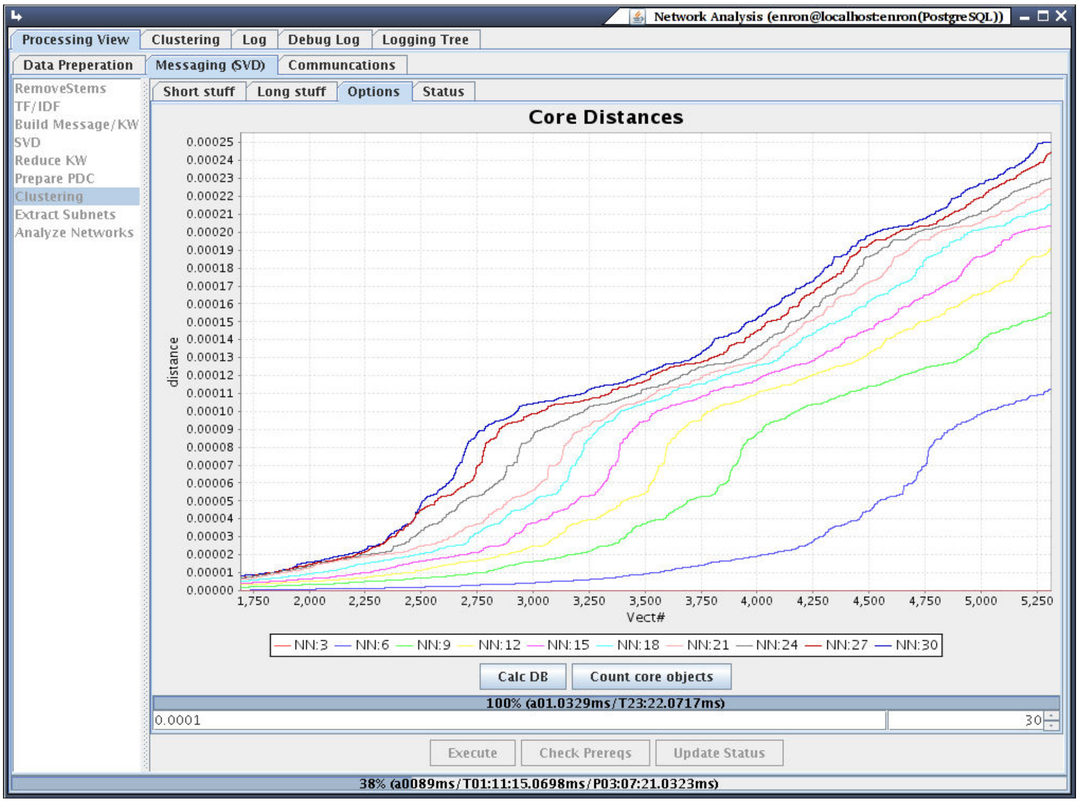
We have implemented a number of clustering approaches, ranging from the classical to the newer approaches from density based data mining:

- **Single-Link:**

  A simple hierarchical divide-and-conquer algorithm [43] used primarily as a reference algorithm.

- **K-Means:**

  This algorithm [85] might be considered an ageing work-horse of data mining, but



100% (a01.0329ms/T23:22.0717ms)

0.0001

30

38% (a0089ms/T01:11:15.0698ms/P03:07:21.0323ms)

Figure 5.7: A core distance distribution

it is surprisingly common in commercial data mining practices today, as well as serving as a standard to measure the performance of other algorithms against.

- **DBScan:**

  The first of two density based clustering algorithms with noise detection. This algorithm [42] performed most of the work during our clustering for topic detection.

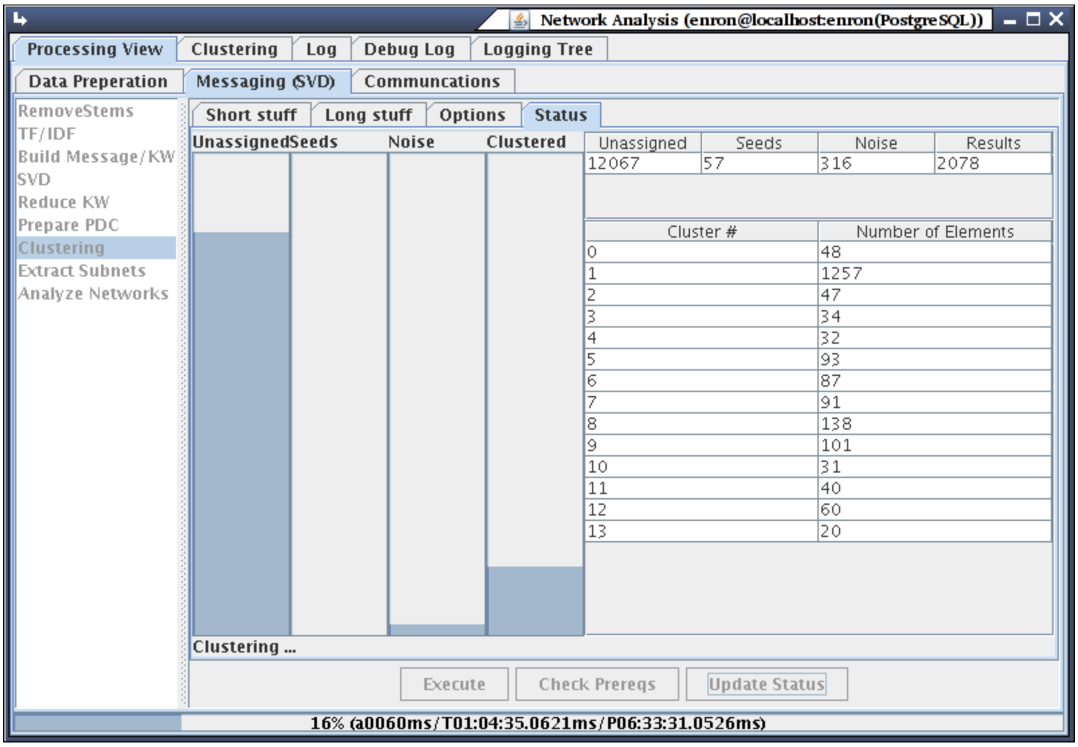
- **OPTICS:**

  The second in the category density based clustering. The preference of the previously mentioned algorithm over this one stems from the fact that OPTICS needs a degree of manual interpretation or control[18, 17, 114].

- **Appleseed:**

  While not involved in the case studies presented in this thesis, this algorithm nevertheless was regarded as an aspect of graph analysis used during the course of the work leading up to this thesis[136]. Recommender systems form a related area of analysis and partially rely on social network analysis.

- **Density Based Graph Mining:**



Unassigned	Seeds	Noise	Results
12067	57	316	2078

Cluster #	Number of Elements
0	48
1	1257
2	47
3	34
4	32
5	93
6	87
7	91
8	138
9	101
10	31
11	40
12	60
13	20

16% (a0060ms/T01:04:35.0621ms/P06:33:31.0526ms)

Figure 5.8: An OPTICS clustering run

> The idea proposed in chapter 3.3 was implemented as a wrapper around density based clustering algorithms such as OPTICS or DBScan. Both algorithms have been used to detect dense sub-graphs (see section 4.4.1 on page 74).

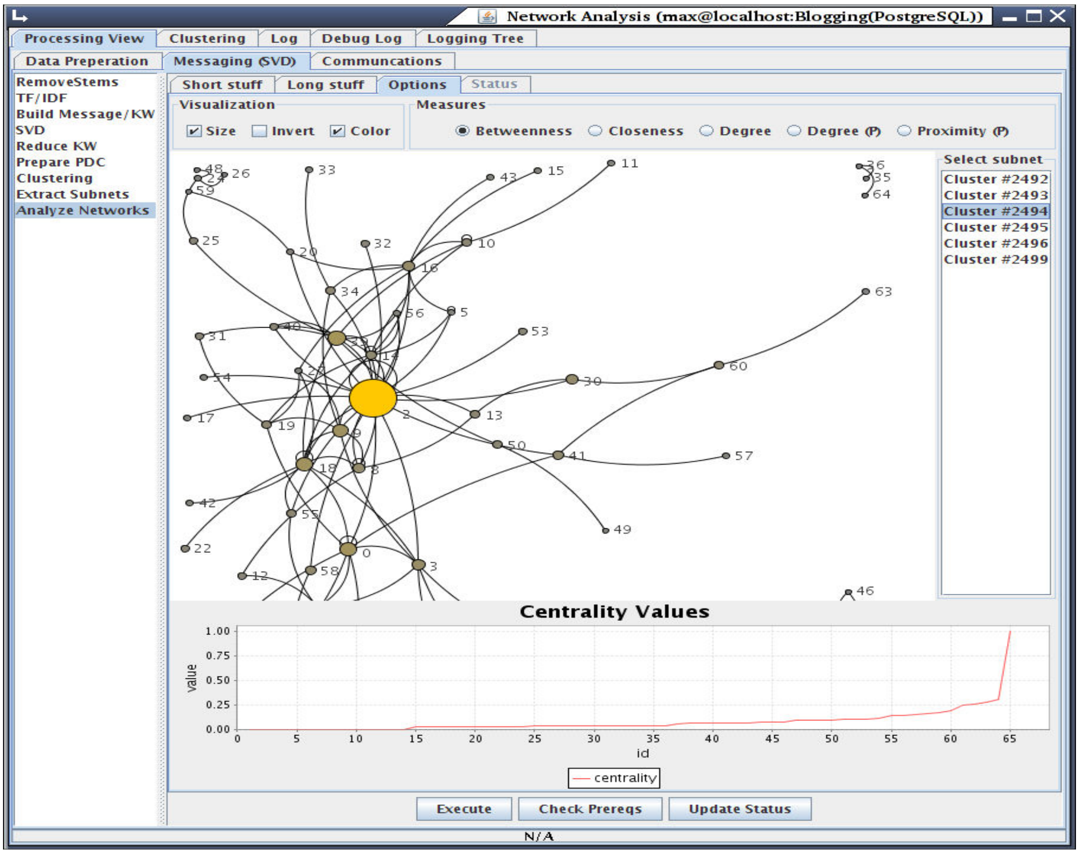
During the course of this work the main algorithm used for topic detection was DBScan. Figure 5.7 shows a nearest neighbor distribution used to gauge where to place the threshold for "core objects" which comprise clusters. Figure 5.8 shows a clustering being performed on a data set, with status bars indicating the current number of elements not yet classified, tagged as possible cluster members, or fully classified as noise or cluster objects.

5.2.4 Social Network Analysis

After topics have been identified, the final analysis can be applied to the now extracted sub-graphs. Figure 5.9 shows a sample extracted sub-graph in the application environment. By being able to switch between the various centrality and prestige models, a quick overview of the data can yield a comprehensive view of the sub-graph at hand.

The implemented centrality and prestige approaches are the following:

- In/Out link centrality



Network Analysis (max@localhost:Blogging(PostgreSQL))

Visualization: ☑ Size ☐ Invert ☑ Color

Measures: ● Betweenness ○ Closeness ○ Degree ○ Degree (P) ○ Proximity (P)

Select subnet: Cluster #2492, Cluster #2493, Cluster #2494, Cluster #2495, Cluster #2496, Cluster #2499

Figure 5.9: A sample extracted sub-network

Basically a measure based on the number of links going in as well as out from an actor. The more links he has, the more connected and embedded within the network he is. While it works well on small networks, it is intrinsically local and takes no account of transitivity in the graph.

- Betweenness centrality

  Probably the most common measure of centrality implemented deems an actor to be central to the network in direct relation to how many paths it lies in between all other pairs of nodes.

- Rank centrality

  Can be taken as an extension of the in/out link centrality idea, only encompassing transitivity. Thus indirect communication to nodes further away in the network are also considered to add to the centrality of an actor.

- In-Link prestige

  Contrary to the measure of centrality mentioned above, this measure aims to capture how many actors refer to a given actor as a measure to how much he is sought after. The more in-links, the more prestige (or perhaps expertize) he has.

- Rank prestige

  Similarly, the determination of prestige under consideration of transitivity yields the Rank prestige measure. As such it is more comprehensive and sensitive than the localized version above.

5.3 Performance and Outlook

Since this thesis provides the groundwork for future automatization and application of partitioning massive graphs for social network analysis, we will take a short look at the performance and efficiency of the techniques presented and used in section 3.3 and chapter 4.

5.3.1 Performance

An important aspect of any analysis is the required time to generate results. It is obviously not the point of this thesis to propose a viable productive solution to partition massive graphs, but rather to pave the way for a future solution. As such it is important to present an overview of the implications for any future analysis.

5.3.2 Related Techniques

A number of steps in the analysis were performed by rather costly steps.

- SVD

  Singular value decomposition was performed within an R environment [110]. Due to the size of the sparse matrices the computation took a rather long time (49h37m22s in the case of Enron). This is a significant hurdle to overcome when just-in-time analyses are to be considered.

- Eigenvalue Analysis

  The calculation of an Eigenspace solution to a given communications matrix is likewise a time intensive project. In our case the processing of a newsgroup took more than three days, which clearly rules out this approach for daily use.

- TF/IDF

  While not as complex as the previous two steps, the method implemented in this thesis rested at the database level, and can be considered to be quite efficient. Nonetheless nearly an entire day was needed to process all keyword vectors in the Enron database.

- Clustering

  The clustering approaches tested in this thesis performed in line with the expectations derived from their complexity. K-Means performed rather adroitly, but did not yield high quality results, whereas DBScan took nearly one and a half days to finish, but resulted in use-able topics.

Any endeavour extending this approach will have to clear most of these hurdles, and at least in a few cases the answers lie in the direction of incremental updates. Since this reaches beyond the scope of this thesis it is considered future work.

5.3.3 Synthetic Data

Any implemented approach was subjected to testing on synthetic, thus known, data. The particular methods for constructing test sets has necessitated the investigation of appropriate ways how to build it. Since we look at graph data as well as vector spaces, both approaches have been implemented.

Simulating Random Graphs

The use of generating random networks has been shown in any field concerned with complex network analysis, such as social network analysis.

We have concentrated on a few basic network generation practices:

- Random Connection

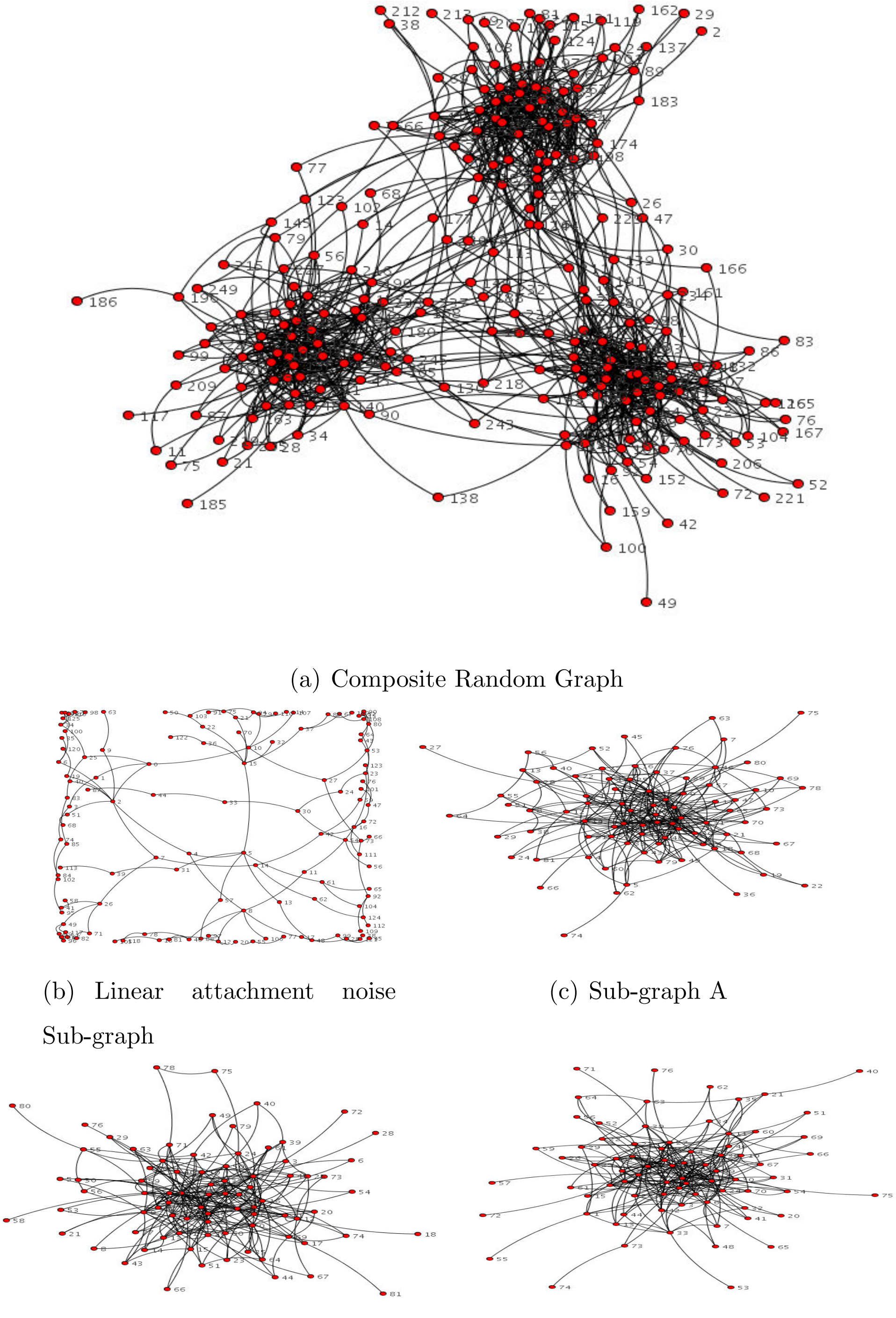
  Concerned with building a sample network between a given set of actors, this method is the most straightforward of the approaches mentioned here. Basically a random number larger than zero yet smaller than the maximal number of edges from a graph are placed by randomly selecting two nodes for each edge. Since the likelihood of each connection is the same, this approach is used to generate the equivalent of background noise.

- Preferential Attachment [13, 11]

  A very effective method to generate simple communication networks. By increasing the chance of an actor gaining an edge with increasing number of exiting connections the basic idea of a social network can be approximated rather well. The resulting graph does not satisfy but rather approximates the scale-free requirement of a communications network.

- R-Mat [25]

  An algorithmic approach to creating scale-free graphs. While the graphs display the properties required, this algorithm was considered beyond the scope of implementation for this thesis. The results obtained from preferential attachment fulfilled all the necessary requirements posed by a communications network, and the need for this additional approach was regarded as small.



(a) Composite Random Graph

(b) Linear attachment noise Sub-graph

(c) Sub-graph A

(d) Sub-graph B

(e) Sub-graph C

Figure 5.10: Random Graphs

Synthetic Vector-space Data

The second aspect to data generation was vector space data. Since the data is geared towards clustering, the data must contain noise as well as more or less clearly defined regular and irregular placed clusters of data. The data generated is constrained by a minimum and maximum, thus creating a bounded vector space.

A sample constructed data set can be seen in Figure 5.11(a). There are five classes of data constructed:

- Noise

  Homogeneous data placed throughout the vector space under consideration (Figure 5.11(b)).

- Spherical Cluster

  A data set centered around a given point in space, with the distance to this point determined by a Gaussian distribution (see Figure 5.11(c)).

- Elliptical Cluster

  Unlike the spherical cluster this cluster type is stretched along several axes randomly through the vector space. A sample can be seen in Figure 5.11(e) in the form of an oblong cluster. In this case two random axes in space were chosen to stretch a spherical cluster into the displayed ellipse.

- Toroid Cluster

  By having a given distance on a plane from a central point in space and calculating a random distance and offset from this mean distance from the center a torus of vectors can be created. The distances are also distributed in a Gaussian fashion.

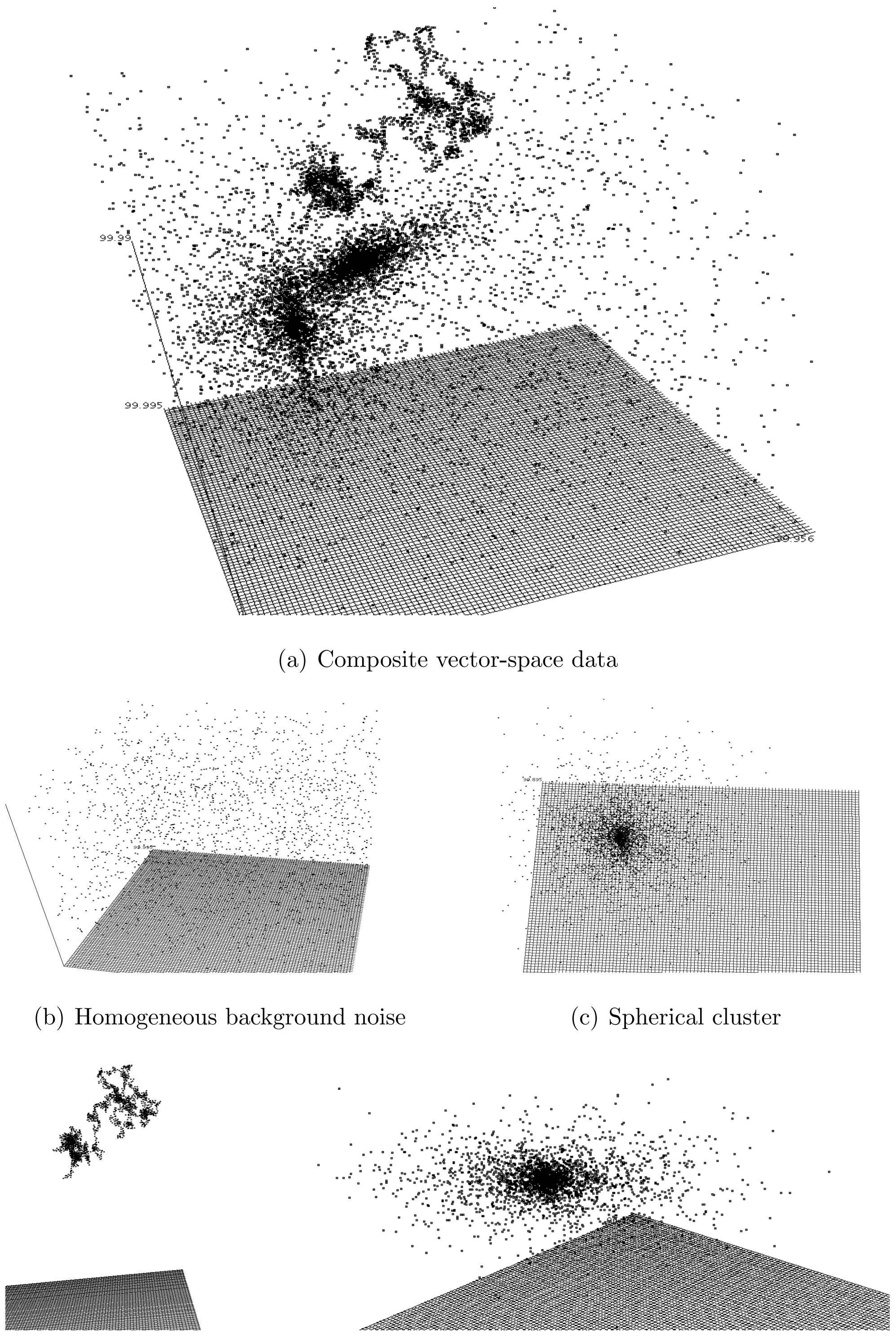
- Blob Cluster

  By selecting a random starting point, and performing a random walk of varying step length and direction through the space a compact but irregular blob of vectors is created (Figure 5.11(d)).

5.3.4 Outlook

The implementation of the algorithms and user interfaces which have made the analysis possible have been to a large degree performed with an eye to better understand the details of the data mining and social network analysis process. To this end it is expected for the framework to be expanded and amended in the future, always being guided by the future work and research to be done in this area of research.

But precisely because the goal of this implementation has been the academic understanding, it is not planned to introduce this software suite into commercial use. While



(a) Composite vector-space data

(b) Homogeneous background noise

(c) Spherical cluster

(d) Blob cluster

(e) Oblong cluster

Figure 5.11: Three dimensional synthetic vector-space data

parts of the analysis framework presented in this thesis will continue to fulfill an important role in preliminary research, especially when considering social network analysis, the functionality with regards to other areas of research will most likely be included from additional sources.

Future Work

A number of additions, next to the completion of supplied documentation, are planned, such as the expansion of network analysis aspects, the generalization of database interactivity and the inclusion of further text mining functionality are planned.

Last but not least the extension of visualisation and graphical representation of networks as well as text corpora is also planned, as the framework provided by this thesis will provide the backdrop for further research into the areas of text mining, social network analysis and graph mining.

6 Conclusion

The application of social network analysis to graphs found in the World Wide Web and the Internet has received increasing attention in recent years. Networks as diverse as those generated by e-mail communication, instant messaging, link structure in the Internet as well as citation and collaboration networks have all been treated with this method. So far these analyses solely utilize graph structure. There is, however, another source of information available in messaging corpora, namely content.

Modern communication networks have been increasing in scope and activity in recent years. The more people partake in email exchanges, and the larger the Internet communities grow, the more traffic is generated. Networks under scrutiny today are increasing in size and complexity, being often an amalgamation and super-positioning of multiple networks. These massive communications corpora can be captured by communications graphs, which are increasingly difficult to the generalized application of social network analysis.

We propose to apply the field of content analysis to the process of social network analysis. By extracting relevant and cohesive sub-networks from massive graphs, we obtain information on the actors contained in such sub-networks to a much firmer degree than before.

While social network analysis can extract the relevance and impact of actors within a network from structural considerations, it does not regard content of the network under scrutiny. Thus the consistency of the network has a great impact upon the interpretability of the results of social network analysis. But as networks get larger, the more general and imprecise are the conclusions drawn from social network analysis. Google will score a very high centrality, based solely on the fact it links to almost everything. Therefore the intention of the analysis methods does not fall into line with the expected data found in real world problems.

We believe the combination of content with social network analysis will greatly improve our understanding of real-world communication.

6.1 Limits of Social Network Analysis

This thesis has focused on the problems posed by social network analysis of massive communications networks. Not only does the expressiveness of social network analysis decrease with increasing graph size, but the larger and more varied the graph the more ambiguous are the conclusions drawn from analysis of important actors within them.

This thesis touches upon several disparate fields of computer science, on the one hand using graph theory as well as incorporating network analysis with the field of text analysis and text mining. By partitioning massive communication graphs, we are able to instill a more focused meaning upon social network analysis of relevant sub-graphs. This has been followed up in two main directions, on the one hand by extracting sub-networks determined by common topics and possibly tracking such topic evolution over time, and on the other hand by focusing on the structure of the network and extracting network leaders or dense communication sub-networks.

6.2 Integrating Content into Social Network Analysis

We have taken the inherent assumption of **social network analysis**, namely the disregard of content when judging **centrality** and **prestige** of actors in a social network, into consideration when analyzing massive graphs. The idea of calculating the centrality of a large, **semantically not homogeneous network**, dilutes the expressiveness of social network analysis to the degree of generality. This has lead us to take content into account when performing social network analysis. By focusing on the topics contained in communication networks, we can pull **topic based sub-networks** out of the massive graph.

Such a topic based network, when treated to social network analysis, now describes centrality and prestige in a much more precise setting than the entire graph. This approach has been presented in the chapters 3.3 through 3.5, and has been used to perform analysis on real world data, namely the Enron data corpus in chapter 4.1 and other text-based communication corpora gleaned from weblogs and news group environments as presented in the chapters 4.2 and 4.3.

In our case we find that a broad content cluster concerned with office humour tends to be spread not by the management, but rather by actors found lower in the corporate hierarchy. This is by no means a clear-cut rule, as even a CEO may find a joke worth sending on. By joining the areas of topic analysis and social network analysis, we can now segment a communication corpus based on text messages into topics, each of which can be analyzed separately to yield **topic based centrality** values. Whereas the result is not as pronounced as in the Enron data set, we have found relevant cohesive sub-networks in both cases. The resulting increased expressiveness of topic based centrality has consequently yielded a better understanding of the communications taking place within the complete communication networks.

6.2.1 Partitioning Massive Graphs

Discovery and tracking of topics in text corpora has previously focused on dissecting the time period into sliding windows of arbitrary size, and subsequently guessing at the correct window size and overlap which critically impact the recognition of topics in their life-cycle. Thus the topics recognized are greatly dependant upon the selection of parameters defining window size and extent. We have moved beyond this problem with a multi-tiered temporal topic tracking approach. The topics can then be used to locate important actors and track their importance over time.

Discovery and tracking of topics in text corpora has previously focused on dissecting the time period into sliding windows of arbitrary size, and subsequently guessing at the correct window size and overlap which critically impact the recognition of topics in their life-cycle. Thus the topics recognized are greatly dependant upon the selection of parameters defining window size and extent.

We developed an approach utilizing several tiers of sliding windows in order to capture topics of varying longevity. Each tier clusters topics at a different level, either short term topics or long term trends. By characterizing each topic level using representative keyword vectors, links between the tiers can be established. Each topic contains a word distribution characteristic of said topic, this distribution is captured as a *foreground model*. Each foreground model can now be contrasted to the word distribution in the entire corpus, described by the *background model*.

In order to compare topics across differing time frames with respect to the overall text corpora, we are able to use **representative keyword vectors** to sum up topics by contrasting the topics occurring within a foreground model with the general background model. Representative keyword vectors are established by discovering salient terms within a topic, such that common topic terms are uncommon within the text corpus in general.

By introducing a multi-tier temporal hierarchy in chapter 3.6, we have been able to follow topic trends as the evolution of long term topic clusters, expressed by the emergence and disappearance of short-term topics linked to the long term trend.

We have implemented and tested this approach on real world data concerning commercial news providers. based on this test phase, the proposed approach will enter test trials at Siemens Corporate Technology. We intend to have a prototype system up and running in the near future.

6.2.2 Interpreting Sub-graphs

Furthermore the use of Eigenvector based social network analysis can also lead to the discovery of relevant topics and central actors. This approach has been followed in chapter 4.2 and implemented in section 4.2.2. These dynamics would have applications for example in detecting shifts of interest or the rise of new 'buzz words' in a given community. Thus the 'spin-doctors' and the 'pet-topics' would become visible.This could be relevant for example for the development of innovative products and services if such an analysis could be performed to 'keep an eye' on topics in R&D communities.

As a conclusion it can be stated that the method proposed can yield interesting insights into the 'mechanics' and 'dynamics' of communicating groups as could be shown in this feasibility study. To gain valid insights from the combination of the analysis of the communication behavior with content analysis it is necessary to perform a considerable effort in data preprocessing such as natural language processing within a concept tailored to the goals of the analysis and the linguistic specifics of the text corpus. Especially the dynamics of actor-to-actor communication and the dynamics of use-of-words make this necessary but worthwhile.

6.2.3 Analyzing Real World Data

The implementation of the algorithms and user interfaces which have made the analysis possible have been to a large degree performed with an eye to better understand the details of the data mining and social network analysis process. To this end it is expected for the framework to be expanded and amended in the future, always being guided by the future work and research to be done in this area of research.

But precisely because the goal of this implementation has been the academic understanding, it is not planned to introduce this software suite into commercial use. While parts of the analysis framework presented in this thesis will continue to fulfill an important role in preliminary research, especially when considering social network analysis, the functionality with regards to other areas of research will most likely be included from additional sources.

The data sets in question were quite large and complex, but every set of real world data presents its own caveats. We ran into numerous instances which necessitated particular tuning of our approach. This indicates a great deal of exploration of the quirks must be made if a large-scale application should be developed from these approaches.

Text mining, specifically the use of keywords as a basis for a vector space representation of documents is still more an art than a science. Particularly the prevalence of informal spelling in short textual messages presents a significant impediment to the use of automated keyword extraction techniques. While the use of dictionaries and possible words within a short Levenshtein distance can correct numerous spelling errors, the use of idioms and colloquial terms increases the difficulty of parsing and analyzing text without human supervision.

The Enron data set also presented a pervasive problem faced in real world data, namely that of incompleteness. As was noted in the section 4.1 the data present is most likely a collection of in-boxes found on email servers at the time of the investigation. This data most likely represents only a part of the email network, which must be captured at the source if the demand for completeness is to be satisfied.

6.3 Results

Our approach has shown the viability of partitioning massive communications corpora. Two real world data sets have undergone the partitioning process: firstly the Enron data set (an e-mail network) and secondly a blogging network. This thesis has also shed some light on the formidable obstacles to the timely analysis of massive corpora, as several analyses are tied to time intensive processing steps.

6.4 Future Work

There are numerous avenues of exploration opened up during the course of this thesis. Apart from the computational challenges presented by the analysis of the partitioning process itself, a number of questions in related fields have been touched upon.

Text mining, specifically the use of keywords as a basis for a vector space representation of documents is still more an art than a science. Particularly the prevalence of informal spelling in short textual messages presents a significant impediment to the use of automated keyword extraction techniques. While the use of dictionaries and possible words within a short Levenshtein distance can correct numerous spelling errors, the use of idioms and colloquial terms increases the difficulty of parsing and analyzing text without human supervision.

List of Tables

List of Figures

Bibliography

[1] L. Adamic and E. Adar. How to search a social network. *Social Networks*, 27(3):187–203, 2005.

[2] L. Adamic, R. Lukose, A. Puniyani, and B. Huberman. Search in power-law networks. *Physical Review E*, 64(4):46135, 2001.

[3] R. Agrawal, S. Rajagopalan, R. Srikant, and Y. Xu. Mining newsgroups using networks arising from social behavior. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pages 529–535, New York, NY, USA, 2003. ACM Press.

[4] A. Aizawa. An information-theoretic perspective of tf-idf measures. *Information Processing and Management*, 39(1):45–65, 2003.

[5] R. Albert, H. Jeong, and A. Barabasi. Diameter of the World-Wide Web. *Nature(London)*, 401(6749):130–131, 1999.

[6] R. Albert, H. Jeong, and A. Barabasi. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382, 2000.

[7] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study: Final report. *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[8] B. Amento, L. Terveen, W. Hill, D. Hix, and R. Sschulman. Experiments in social data mining: The topicshop system. *ACM Transactions on Computer-Human Interaction*, 10(1):54–85, 2003.

[9] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander. Optics: Ordering points to identify the clustering structure. In *SIGMOD Conference*, pages 49–60, 1999.

[10] J. Balthrop, S. Forrest, M. Newman, and M. Williamson. Technological networks and the spread of computer viruses. *Science*, 304(5670):527–529, 2004.

[11] A. Barabasi. *Linked: how everything is connected to everything else and what it means for business, science, and everyday life.* Plume Books, 2003.

[12] A. Barabasi and E. Bonabeau. Scale-free networks. *Scientific American*, 288(5):60–9, 2003.

[13] A. Barabasi and R. Crandall. *Linked: The New Science of Networks*, volume 71. American Journal of Physics, 2003.

[14] B. Bollobás. *Modern Graph Theory.* Springer, 1998.

[15] A. Borodin, G. O. Roberts, J. S. Rosenthal, and P. Tsaparas. Finding authorities and hubs from link structures on the world wide web. *WWW '01: Proceedings of the 10th international conference on World Wide Web*, pages 415–429, 2001.

[16] D. Boyd and J. Potter. Social network fragments: an interactive tool for exploring digital social connections. In *GRAPH '03: Proceedings of the SIGGRAPH 2003 conference on Sketches & applications*, pages 1–1, New York, NY, USA, 2003. ACM Press.

[17] S. Brecheisen, H. Kriegel, P. Kroger, M. Pfeifle, M. Viermetz, and M. Potke. Boss: browsing optics-plots for similarity search. *Proceedings. 20th International Conference on Data Engineering*, page 858, 2004.

[18] S. Brecheisen, H.-P. Kriegel, P. Kroeger, M. Pfeifle, and M. Viermetz. Representatives for visually analyzing cluster hierarchies. In *Proc. 4th Int. Workshop on Multimedia Data Mining (MDM/KDD'03), Washington DC*, pages 64–71, 2003.

[19] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *WWW7: Proceedings of the seventh international conference on World Wide Web 7*, pages 107–117, Amsterdam, The Netherlands, The Netherlands, 1998. Elsevier Science Publishers B. V.

[20] B. Brown and M. Bell. CSCW at play: 'there' as a collaborative virtual environment. In *CSCW '04: Proceedings of the 2004 ACM conference on Computer supported cooperative work*, pages 350–359, New York, NY, USA, 2004. ACM Press.

[21] R. Burt. Positions in networks. *Social Forces*, 55(1):93–122, 1976.

[22] O. Büyükkökten. Orkut: A social network website. http://www.orkut.com, 2007.

[23] CERN - European Organization for Nuclear Research. Colt linear algebra package. http://acs.lbl.gov/∼hoschek/colt/, 2007.

[24] D. Chakrabarti and C. Faloutsos. Graph mining: Laws, generators, and algorithms. *ACM Comput. Surv.*, 38(1):2, 2006.

[25] D. Chakrabarti, Y. Zhan, and C. Faloutsos. R-mat: A recursive model for graph mining. April 2004.

[26] S. Chakrabarti, M. M. Joshi, K. Punera, and D. M. Pennock. The structure of broad topics on the web. In *WWW '02: Proceedings of the 11th international conference on World Wide Web*, pages 251–262, New York, NY, USA, 2002. ACM Press.

[27] J. Chart. Jfreechart: A free java chart library. http://www.jfree.org/index.html, 2007.

[28] C. Chen. The centrality of pivotal points in the evolution of scientific networks. *Proceedings of the 10th international conference on Intelligent user interfaces*, pages 98–105, 2005.

[29] P. Chirita, A. Damian, W. Nejdl, and W. Siberski. Search strategies for scientific collaboration networks. *Proceedings of the 2005 ACM workshop on Information retrieval in peer-to-peer networks*, pages 33–40, 2005.

[30] W. W. Cohen. Enron email dataset. http://www-2.cs.cmu.edu/∼enron/, 2006.

[31] D. Cook and L. Holder. Substructure Discovery Using Minimum Description Length and Background Knowledge. *Journal of Artificial IntelligenceResearch*, 1:231–255, 1994.

[32] D. Cook and L. Holder. *Mining Graph Data.* John Wiley & Sons, 2006.

[33] A. Culotta, R. Bekkerman, and A. McCallum. Extracting social networks and contact information from email and the web. *Proceedings of CEAS, First Conference on Email and Anti-Spam (CEAS)*, 2004.

[34] A. Dasgupta, R. Kumar, P. Raghavan, and A. Tomkins. Variable latent semantic indexing. In *KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 13–21, New York, NY, USA, 2005. ACM Press.

[35] I. Dhillon, Y. Guan, and B. Kulis. A fast kernel-based multilevel algorithm for graph clustering. In *KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 629–634, New York, NY, USA, 2005. ACM Press.

[36] P. Domingos and M. Richardson. Mining the network value of customers. In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 57–66, New York, NY, USA, 2001. ACM Press.

[37] P. Dourish, W. K. Edwards, A. LaMarca, J. Lamping, K. Petersen, M. Salisbury, D. B. Terry, and J. Thornton. Extending document management systems with user-specific active properties. *ACM Trans. Inf. Syst.*, 18(2):140–170, 2000.

[38] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993.

[39] P. Erdös and A. Renyi. On random graphs. *Publicationes Mathematicae*, 6:290–297, 1959.

[40] P. Erdös and A. Renyi. On the evolution of random graphs. *Bulletin of the Institute of International Statistics*, 5:17–61, 1960.

[41] T. Erickson and W. Kellogg. Social translucence: an approach to designing systems that support social processes. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 7(1):59–83, 2000.

[42] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD'06*, pages 226–231, 1996.

[43] M. Ester and J. Sander. *Knowledge Discovery in Databases*. Springer, 2000.

[44] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery. *Advances in Knowledge Discovery and Data Mining, Melno Park*, pages 1–24, 1996.

[45] G. W. Flake, S. Lawrence, and C. L. Giles. Efficient identification of web communities. In *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–160, New York, NY, USA, 2000. ACM Press.

[46] J. Fowler and S. Jeon. The authority of Supreme Court precedent. *Social Networks*, 30(1):16–30, 2008.

[47] P. Fraigniaud, C. Gavoille, and C. Paul. Eclecticism shrinks even small worlds. *Distributed Computing*, 18(4):279–291, 2006.

[48] L. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, 1977.

[49] L. Freeman. Centrality in social networks: Conceptual clarification. *Social Networks*, 1(3):215–239, 1979.

[50] L. Freeman. Finding social groups: A meta-analysis of the southern women data. *Dynamic Social Network Modeling and Analysis. The Natl Acad. Press, Washington, DC*, 2003.

[51] S. Freeman and L. Freeman. *The Networkers Network: A Study of the Impact of a New Communications Medium on Sociometric Structure*. School of Social Sciences Univ. of Calif, 1979.

[52] J. Friedman, F. Baskett, and L. Shustek. An algorithm for finding nearest neighbors. *IEEE Transactions on Computers*, 24(10):1000–1006, 1975.

[53] D. Gibson, R. Kumar, and A. Tomkins. Discovering large dense subgraphs in massive graphs. *Proceedings of the 31st international conference on Very Large Data Bases*, pages 721–732, 2005.

[54] J. Goecks and E. D. Mynatt. Leveraging social networks for information sharing. In *CSCW '04: Proceedings of the 2004 ACM conference on Computer Supported Cooperative Work*, pages 328–331, New York, NY, USA, 2004. ACM Press.

[55] G. Golub and C. Reinsch. Singular value decomposition and least squares solutions. *Numerische Mathematik*, 14(5):403–420, 1970.

[56] M. S. Granovetter. The strength of weak ties. *The American Journal of Sociology*, 78(6):1360–1380, 1973.

[57] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins. The predictive power of online chatter. In *KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 78–87, New York, NY, USA, 2005. ACM Press.

[58] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. *Proceedings of the 13th international conference on World Wide Web*, pages 491–501, 2004.

[59] T. H. Haveliwala. Topic-sensitive pagerank. In *WWW '02: Proceedings of the 11th international conference on World Wide Web*, pages 517–526, New York, NY, USA, 2002. ACM Press.

[60] J. Hicklin, C. Moler, P. Webb, R. F. Boisvert, B. Miller, R. Pozo, and K. Remington. Jama : A java matrix package. http://math.nist.gov/javanumerics/jama/#Package, 2006.

[61] L. Hinrichs. Xing: The small world of business professionals. http://www.xing.com/, 2006.

[62] P. Hoff, A. Raftery, and M. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1099, 2002.

[63] R. Hoffman and D. Nye. Linkedin: A social network website. http://www.linkedin.com/, 2006.

[64] T. Hofmann. Probabilistic latent semantic indexing. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, 1999.

[65] L. Holder, D. Cook, and S. Djoko. Substructure discovery in the subdue system. *Proc. AAAI*, 94:169–180, 1994.

[66] J. Hopcroft, O. Khan, B. Kulis, and B. Selman. Natural communities in large linked networks. *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pages 541–546, 2003.

[67] D. B. Horn, T. A. Finholt, J. P. Birnholtz, D. Motwani, and S. Jayaraman. Six degrees of jonathan grudin: a social network analysis of the evolution and impact of cscw research. In *CSCW '04: Proceedings of the 2004 ACM conference on Computer supported cooperative work*, pages 582–591, New York, NY, USA, 2004. ACM Press.

[68] B. Hoser. *Analysis of Asymmetric Communication Patterns in Computer Mediated Communication Environments.* PhD thesis, Universität Karlsruhe, 2004.

[69] B. Hoser, J. Schröder, A. Geyer-Schulz, M. Viermetz, and M. Skubacz. Topic trend detection in newsgroups. *Kuenstliche Intelligenz*, Special Issue, 2007.

[70] H. Ino, M. Kudo, and A. Nakamura. Partitioning of web graphs by community topology. *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 661–669, 2005.

[71] A. Jain and R. Dubes. *Algorithms for Clustering Data.* Prentice-Hall advanced reference series. Prentice-Hall, Inc, Upper Saddle River, NJ., 1988.

[72] I. Jonyer, D. J. Cook, and L. B. Holder. Graph-based hierarchical conceptual clustering. *J. Mach. Learn. Res.*, 2:19–43, 2002.

[73] H. Kashima and A. Inokuchi. Kernels for graph classification. *ICDM Workshop on Active Mining*, 2002.

[74] N. Katayama and S. Satoh. The sr-tree: An index structure for high-dimensional nearest neighbor queries. In J. Peckham, editor, *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data, May 13-15, 1997, Tucson, Arizona, USA*, pages 369–380. ACM Press, 1997.

[75] D. Kempe, J. Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146, New York, NY, USA, 2003. ACM Press.

[76] J. Kleinberg. The small-world phenomenon: an algorithm perspective. In *STOC '00: Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 163–170, New York, NY, USA, 2000. ACM Press.

[77] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In *Journal of the ACM*, volume 46, pages 604–632, New York, NY, USA, 1999. ACM Press.

[78] V. E. Krebs. Mapping networks of terrorist cells. *Connections*, 24(3):43–52, 2002.

[79] S. Kullback and R. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[80] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. *World Wide Web*, 8(2):159–178, 2005.

[81] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tompkins, and E. Upfal. The web as a graph. *Proceedings of the nineteenth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 1–10, 2000.

[82] D. Liben-Nowell. *An Algorithmic Approach to Social Networks*. PhD thesis, Massachusetts Institute of technology, 2005.

[83] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, pages 556–559, New York, NY, USA, 2003. ACM Press.

[84] D. Lin and P. Pantel. Concept discovery from text. *Proceedings of the 19th international conference on Computational linguistics*, 1:1–7, 2002.

[85] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Prob.*, volume Volume 1, pages 281–297, 1967.

[86] P. Marsden. Methods for the characterization of role structures in network analysis. *Research Methods in Social Network Analysis*, pages 489–530, 1989.

[87] N. Matsumura, D. E. Goldberg, and X. Llorà. Mining directed social network from message board. In *WWW '05: Special interest tracks and posters of the 14th international conference on World Wide Web*, pages 1092–1093, New York, NY, USA, 2005. ACM Press.

[88] M. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409, 2001.

[89] M. Newman. A measure of betweenness centrality based on random walks. *Social Networks*, 27(1):39–54, 2005.

[90] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

[91] M. E. J. Newman. Detecting community structure in networks. *The European Physical Journal B*, 38:321 – 330, 2004.

[92] C. Noble and D. Cook. Graph-based anomaly detection. *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 631–636, 2003.

[93] J. O'Madadhain, D. Fisher, S. White, and Y. Boey. The JUNG (java universal network/graph) framework. Technical report, Technical report, UC Irvine, 2003.

[94] J. On. They rule: A social network application of corporate america. http://www.theyrule.net, 2007.

[95] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Brining order to the web. Technical report, Stanford Univerity, Computer Science Department, 1998.

[96] G. Penchina. Wikipedia, the online encyclopedia. http://www.wikipedia.org/, 2005.

[97] J. S. Quarterman and J. C. Hoskins. Notable computer networks. In *Communications of the ACM*, volume 29, pages 932–971, New York, NY, USA, 1986. ACM Press.

[98] P. Reynolds. The oracle of bacon. http://oracleofbacon.org/, 2005.

[99] M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 61–70, New York, NY, USA, 2002. ACM Press.

[100] M. F. Schwartz and D. C. M. Wood. Discovering shared interests using graph analysis. In *Commununications of the ACM*, volume 36, pages 78–89, New York, NY, USA, 1993. ACM Press.

[101] R. Sibson. SLINK: An optimally efficient algorithm for the single-link cluster method. *The Computer Journal*, 16(1):30–34, 1973.

[102] X. Song, C.-Y. Lin, B. L. Tseng, and M.-T. Sun. Modeling and predicting personal information dissemination behavior. In *KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 479–488, New York, NY, USA, 2005. ACM Press.

[103] E. Spertus, M. Sahami, and O. Buyukkokten. Evaluating similarity measures: a large-scale study in the orkut social network. In *KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 678–684, New York, NY, USA, 2005. ACM Press.

[104] M. Stefanone, J. Hancock, G. Gay, and A. Ingraffea. Emergent networks, locus of control, and the pursuit of social capital. In *CSCW '04: Proceedings of the 2004 ACM conference on Computer supported cooperative work*, pages 592–595, New York, NY, USA, 2004. ACM Press.

[105] K. Stephenson and M. Zelen. Rethinking centrality: methods and examples. *Social Networks*, 11(1):1–37, 1989.

[106] C. Stolz, M. Viermetz, M. Skubacz, and R. Neuneier. Guidance Performance Indicator - Web Metrics for Information Driven Web Sites. *IEEE Intl. Conf. Web Intelligence 2005, Proc.*, pages 186–192, 2005.

[107] C. Stolz, M. Viermetz, M. Skubacz, and R. Neuneier. Improving semantic consistency of web sites by quantifying user intent. *Springer LNCS: Int. Conf on Web Engineering, ICWE 2005, Sydney*, pages 308–317, 2005.

[108] S. Strogatz. Exploring complex networks. *Nature*, 410:268–276, 2001.

[109] J. Tang, H. Li, Y. Cao, and Z. Tang. Email data cleaning. In *KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 489–498, New York, NY, USA, 2005. ACM Press.

[110] R. D. C. Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2007.

[111] P. Tennent, M. Hall, B. Brown, M. Chalmers, and S. Sherwood. Three applications for mobile epidemic algorithms. *Proceedings of the 7th international conference on Human computer interaction with mobile devices & services*, pages 223–226, 2005.

[112] T. Tomokiyo and M. Hurst. A language model approach to keyphrase extraction. In *Proceedings of the ACL Workshop on Multiword Expressions*, pages 33–40, 2003.

[113] J. Travers and S. Milgram. An experimental study of the small world problem. *Sociometry*, 32(4):425–443, 1969.

[114] M. Viermetz. Boss: Browsing optics plots for similarity search. Master's thesis, Ludwig-Maximilians Universität München, 2003.

[115] M. Viermetz and M. Skubacz. Using topic discovery to segment large communication graphs for social network analysis. IEEE Intl. Conf. Web Intelligence 2007, Proc., 2007.

[116] M. Viermetz, M. Skubacz, C.-N. Ziegler, D. Dittert, and D. Seipel. Tracking topic evolution in news environments. *CEC EEE 2008*, 2007.

[117] M. Viermetz, C. Stolz, M. Barth, and K. Wilde. Searchstrings revealing user intent - a better understanding of user perception. *Springer LNCS: Int. Conf on Web Engineering, ICWE 2006, Palo Alto*, pages 225–232, 2006.

[118] M. Viermetz, C. Stolz, V. Gedov, and M. Skubacz. Relevance and impact of tabbed browsing behavior on web usage mining. *IEEE Intl. Conf. Web Intelligence 2006, Proc.*, pages 262–269, 2006.

[119] S. Wan, S. Wong, and P. Prusinkiewicz. An algorithm for multidimensional data clustering. *ACM Transactions on Mathematical Software (TOMS)*, 14(2):153–162, 1988.

[120] T. Washio and H. Motoda. State of the art of graph-based data mining. In *SIGKDD Explorer Newsletter*, volume 5, pages 59–68, New York, NY, USA, 2003. ACM Press.

[121] S. Wassermann and K. Faust. *Social Network Analysis*. Cambridge University Press, 1994.

[122] D. Watts and S. Strogatz. Collective dynamics of'small-world'networks. *Nature*, 393(6684):409–10, 1998.

[123] D. Welch, N. Buchheit, and A. Ruocco. Strike back: Offensive actions in information warfare. *ACM New Security Paradigm Workshop*, pages 47–52, 1999.

[124] B. Wellman. For a social network analysis of computer networks: a sociological perspective on collaborative work and virtual community. *SIGCPR '96: Proceedings of the 1996 ACM SIGCPR/SIGMIS conference on Computer personnel research*, pages 1–11, 1996.

[125] B. Wellman. A computer network is a social network. In *SIGGROUP Bulletin*, volume 19, pages 41–45, New York, NY, USA, 1998. ACM Press.

[126] H. White. Search parameters for the small world problem. *Social Forces*, 49(2):259–264, 1970.

[127] H. White, S. Boorman, and R. Breiger. Social structure from multiple networks: Blockmodels of roles and positions. *The American Journal of Sociology*, 81(4):730–780, 1976.

[128] S. White and P. Smyth. Algorithms for estimating relative importance in networks. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 266–275, New York, NY, USA, 2003. ACM Press.

[129] S. Wolfram. *Mathematica: a system for doing mathematics by computer*. Addison Wesley Longman Publishing Co., Inc. Redwood City, CA, USA, 1991.

[130] A. Y. Wu, M. Garland, and J. Han. Mining scale-free networks using geodesic clustering. *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 719–724, 2004.

[131] J. J. Xu and H. Chen. Crimenet explorer: a framework for criminal network knowledge discovery. *ACM Trans. Inf. Syst.*, 23(2):201–226, 2005.

[132] J. Yee, R. Mills, G. Peterson, and S. Bartczak. Automatic generation of social network data from electronic-mail communications. *10th ICCRTS*, 2005.

[133] B. Yu and M. P. Singh. Searching social networks. In *AAMAS '03: Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pages 65–72, New York, NY, USA, 2003. ACM Press.

[134] S. Zanero and S. M. Savaresi. Unsupervised learning techniques for an intrusion detection system. *SAC '04: Proceedings of the 2004 ACM symposium on Applied computing*, pages 412–419, 2004.

[135] J. Zhang and M. S. Ackerman. Searching for expertise in social networks: a simulation of potential strategies. *GROUP '05: Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*, pages 71–80, 2005.

[136] C. Ziegler and G. Lausen. Propagation models for trust and distrust in social networks. *Information Systems Frontiers*, 7(4):337–358, 2005.

[137] C. Ziegler, M. Skubacz, A. Siemens, and C. IC. Towards automated reputation and brand monitoring on the web. *Web Intelligence, 2006. WI 2006. IEEE/WIC/ACM International Conference on*, pages 1066–1072, 2006.

[138] C.-N. Ziegler. *Towards Decentralized Recommender Systems.* PhD thesis, Albert-Ludwigs-Universität Freiburg, Freiburg i.Br., Germany, June 2005.

Index