ISCTE IUL
Instituto Universitário de Lisboa
Departamento de Ciências e Tecnologias da Informação

FACULDADE · DE · CIÊNCIAS UNIVERSIDADE · DE · LISBOA
Departamento de Informática

Reading the News Through its Structure: New Hybrid Connectivity Based Approaches

David Manuel de Sousa Rodrigues

Tese especialmente elaborada para a obtenção do grau de doutor
em Ciências da Complexidade
A Thesis presented in partial fulfilment of the Requirements for the Degree of
Doctor of Philosophy in the field of Complexity Sciences

Orientador / Advisor: Professor Jorge Louçã, Professor Auxiliar

November, 2013

ISCTE IUL
Instituto Universitário de Lisboa
Departamento de Ciências e Tecnologias da Informação

FACULDADE · DE · CIÊNCIAS UNIVERSIDADE · DE · LISBOA
Departamento de Informática

Reading the News Through its Structure: New Hybrid Connectivity Based Approaches

David Manuel de Sousa Rodrigues

Composição do Júri / Juri Members

Doutora Tanya Vianna de Araújo, *Universidade de Lisboa, Portugal*
Doutor José Felix Costa, *Universidade de Lisboa, Portugal*
Professor Jeffrey H. Johnson, *The Open University, United Kingdom*
Doutor Helder Manuel Ferreira Coelho, *Universidade de Lisboa, Portugal*
Doutor Rui Jorge Henriques Calado Lopes, *Instituto Universitário de Lisboa, Portugal*
Doutor Jorge Manuel Anacleto Louçã, *Instituto Universitário de Lisboa, Portugal*

Presidido por / Presided by
Doutor Luís Ducla Soares, *Instituto Universitário de Lisboa, Portugal*

Defendida em provas públicas a 17 de Março de 2014

November, 2013

Abstract

In this thesis a solution for the problem of identifying the structure of news published by online newspapers is presented. This problem requires new approaches and algorithms that are capable of dealing with the massive number of online publications in existence (and that will grow in the future). The fact that news documents present a high degree of interconnection makes this an interesting and hard problem to solve. The identification of the structure of the news is accomplished both by descriptive methods that expose the dimensionality of the relations between different news, and by clustering the news into topic groups. To achieve this analysis this integrated whole was studied using different perspectives and approaches.

In the identification of news clusters and structure, and after a preparatory data collection phase, where several online newspapers from different parts of the globe were collected, two newspapers were chosen in particular: the Portuguese daily newspaper **Público** and the British newspaper **The Guardian**.

In the first case, it was shown how information theory (namely variation of information) combined with adaptive networks was able to identify topic clusters in the news published by the Portuguese online newspaper **Público**.

In the second case, the structure of news published by the British newspaper **The Guardian** is revealed through the construction of time series of news clustered by a k-means process. After this approach an unsupervised algorithm, that filters out irrelevant news published online by taking into consideration the connectivity of the news labels entered by the journalists, was developed. This novel hybrid technique is based on Q-analysis for the construction of the filtered network followed by a clustering technique to identify the topical clusters. Presently this work uses a modularity optimisation clustering

technique but this step is general enough that other hybrid approaches can be used without losing generality.

A novel second order swarm intelligence algorithm based on Ant Colony Systems was developed for the travelling salesman problem that is consistently better than the traditional benchmarks. This algorithm is used to construct Hamiltonian paths over the news published using the eccentricity of the different documents as a measure of distance. This approach allows for an easy navigation between published stories that is dependent on the connectivity of the underlying structure.

The results presented in this work show the importance of taking topic detection in large corpora as a multitude of relations and connectivities that are not in a static state. They also influence the way of looking at multi-dimensional ensembles, by showing that the inclusion of the high dimension connectivities gives better results to solving a particular problem as was the case in the clustering problem of the news published online.

Keywords: Adaptive Networks, Q-analysis, Community Detection, Swarm Intelligence, Hamiltonian Path, Travelling Salesman Problem, Ant Colony Optimisation

Resumo

Neste trabalho resolvemos o problema da identificação da estrutura das notícias publicadas em linha por jornais e agências noticiosas. Este problema requer novas abordagens e algoritmos que sejam capazes de lidar com o número crescente de publicações em linha (e que se espera continuam a crescer no futuro). Este facto, juntamente com o elevado grau de interconexão que as notícias apresentam tornam este problema num problema interessante e de difícil resolução. A identificação da estrutura do sistema de notícias foi conseguido quer através da utilização de métodos descritivos que expõem a dimensão das relações existentes entre as diferentes notícias, quer através de algoritmos de agrupamento das mesmas em tópicos. Para atingir este objetivo foi necessário proceder a ao estudo deste sistema complexo sob diferentes perspectivas e abordagens.

Após uma fase preparatória do corpo de dados, onde foram recolhidos diversos jornais publicados online optou-se por dois jornais em particular: O **Público** e o **The Guardian**. A escolha de jornais em línguas diferentes deve-se à vontade de encontrar estratégias de análise que sejam independentes do conhecimento prévio que se tem sobre estes sistemas.

Numa primeira análise é empregada uma abordagem baseada em redes adaptativas e teoria de informação (nomeadamente variação de informação) para identificar tópicos noticiosos que são publicados no jornal português **Público**.

Numa segunda abordagem analisamos a estrutura das notícias publicadas pelo jornal Britânico **The Guardian** através da construção de séries temporais de notícias. Estas foram seguidamente agrupadas através de um processo de k-means. Para além disso desenvolveu-se um algoritmo que permite filtrar de forma não supervisionada notícias irrelevantes que apresentam baixa conectividade às restantes notícias através da utilização de Q-analysis seguida de um processo de *clustering*. Presentemente este método utiliza otimização de

modularidade, mas a técnica é suficientemente geral para que outras abordagens híbridas possam ser utilizadas sem perda de generalidade do método.

Desenvolveu-se ainda um novo algoritmo baseado em sistemas de colónias de formigas para solução do problema do caixeiro viajante que consistentemente apresenta resultados melhores que os tradicionais bancos de testes. Este algoritmo foi aplicado na construção de caminhos Hamiltonianos das notícias publicadas utilizando a excentricidade obtida a partir da conectividade do sistema estudado como medida da distância entre notícias. Esta abordagem permitiu construir um sistema de navegação entre as notícias publicadas que é dependente da conectividade observada na estrutura de notícias encontrada.

Os resultados apresentados neste trabalho mostram a importância de analisar sistemas complexos na sua multitude de relações e conectividades que não são estáticas e que influenciam a forma como tradicionalmente se olha para sistema multi-dimensionais. Mostra-se que a inclusão desta dimensões extra produzem melhores resultados na resolução do problema de identificar a estrutura subjacente a este problema da publicação de notícias em linha.

Palavras Chave: Redes Adaptativas, *Q*-analysis, Detecção de Comunidades, Swarm Intelligence, Caminhos Hamiltonianos, Problema do Caixeiro Viajante, Optimização por Colónias de Formigas

Scientific Publications

- Rodrigues, D. M. S. and Louçã, J. (2009). **Mutual information to assess structural properties in dynamic networks**. In *Proceedings of the European conference on Complex Systems*. Warwick 2009.
- Rodrigues, D. M. S. (2010). **The observatorium – the structure of news: topic monitoring in online media with mutual information**. In Louçã, J., editor, *Proceedings of the European Conference on Complex Systems*. Complex Systems Society, Lisbon.
- Louçã, J. and Rodrigues, D. M. S. (2011). **The observatorium – observation et analyse de reseaux de communication grande echelle**. In *Proceedings of the Rencontres interdisciplinaires sur les systemes complexes naturels et artificiels*, Rochebrune, France.
- Rodrigues, D. M. S., Louçã, J., and Ramos, V. (2011). **From standard to second-order swarm intelligence phase-space maps**. In Thurner, S., editor, *8th European Conference on Complex Systems*, Vienna, Austria.
- Ramos, V., Rodrigues, D. M. S., and Louçã, J. (2011). **Spatio-temporal dynamics on co-evolved stigmergy**. In Thurner, S., editor, *8th European Conference on Complex Systems*, Vienna, Austria.
- Ramos, V., Rodrigues, D. M. S., and Louçã, J. (2013). **Second order swarm intelligence**. In Pan, J.-S., Policarpou, M., Wozniak, M., de Carvalho, A. C., Quintián, H., and Corchado, E., editors, *HAIS'13. 8th International Conference on Hybrid Artificial Intelligence Systems, volume Lecture Notes in Computer Science*, Salamanca, Spain. Springer.
- Rodrigues, D. M. S. (2013). **Identifying news clusters using *Q*-analysis and modularity**. In Albert Diaz-Guilera, Alex Arenas, A. C., editors, In *Proceedings of the European Conference on Complex Systems 2013*, Barcelona.

Acknowledgements

I'd like to express my sincere gratitude to everyone who assisted in the completion of this thesis. The best and worst moments of my doctoral journey have been shared with many. It has been a great privilege to walk this road with them.

I would like to thank my advisor, Professor Jorge Louçã. I am deeply grateful for his incentive, guidance, financial support, and encouragement throughout the years of my stay at ISCTE-IUL.

I am grateful to Professor Jeffrey Johnson for his constructive comments, suggestions, and help in finding some of his and Professor Ron Atkin early Q-analysis texts.

I'm also thankful to all whose paths crossed mine and for all the vivid discussions held on these topics, that helped sharpen my understanding of complex systems. Namely Vitorino Ramos for his brilliant thinking and his scientific integrity; Jane Bromley for always being supportive and encouraging; and David Hales for his ruthless critical thinking.

I would like to thank all my good friends for making the best of the worse times.

My thanks would not be complete without acknowledging the support of my brother Miguel, my sister Eva, my parents David and Conceição, and most importantly my beloved wife Mafalda. Without their unconditional love, faith and patience, this work would never have been a reality.

Contents

List of Figures

List of Tables

List of Algorithms

List of Symbols

Chapter 8

VI	Variation of Information
k	node *k*
t	time step
n	total number of elements in the clustering
$P(k)$	Probability that a node *k* belongs to cluster C_k
C and C'	partitionings of the dataset used to compute *VI* between consecutive states
$H(C)$	Entropy of clustering C
$I(C, C')$	Mutual Information between clustering C and C'
$P(k, k')$	Probability that a node belonging to cluster C_k is in cluster C'_k
$J(A, B)$	Jaccard similarity between two sets A and B
TTL	Time to Live of each node
j_{min}	threshold of the Jaccard index for connecting topics
VI_{min}	threshold of VI for detecting topics

Chapter 9

tf	term frequency
idf	inverse document frequency
tf.idf	term frequency inverse document frequency
u and v	feature vectors representing text documents

Chapter 10

q	connectivity between two simplicies. Two simplicies q-connected share $q+1$ vertices
G_i	Projected Graph from the bipartite graph where connectivity $q \geq i$
μ	modularity

Chapter 11

TSP	Travelling Salesman Problem
ACO	Ant Colony Optimisation
ACS	Ant Colony System
$p_k(r,s)$	probability that at an ant k moves from city r to city s
α	control parameter for the balance between positive and negative pheromone ($\alpha = 1$ is pure ACS)
β	control parameter for the influence of distance between cities in ants choice of path
τ^+	positive pheromone deposited on path
τ^-	negative pheromone deposited on path
η	distance between two cities
ρ^+	evaporation rate of positive pheromone
ρ^-	evaporation rate of negative pheromone
q_0	threshold parameter to balance between exploration and exploitation in ants solution construction
$\Delta\tau^+$ and $\Delta\tau^-$	amount of pheromone to deposit at each update cycle
L_{gb}^{-1}	Inverse of the length of the best tour calculated from a short neighbour heuristic
nL_{gb}^{-1}	Inverse of the length of the worse tour calculated from a far neighbour heuristic
$iter$	iteration
$ecc_{a,b}$	eccentricity between two simplicies a and b

Chapter 1

Introduction

This thesis presents new methods to analyse, in automated ways, news published through online media. The problem this thesis proposes to solve is how to extract the underlying structure of news published by online newspapers, with the least amount of human intervention. In this thesis several questions about this problem are answered, ranging from knowing if there is some structure in the news; if it is possible to find clusters or patterns; and if it is possible to derive methods that automatically extract meaningful information from this great quantity of text. This problem presents some particular difficulties, namely the volume of data produced by newspapers on a daily basis, the gathering of that data, and processing of it in a quick and manageable way.

The main scientific research field of this thesis is network science, more specifically it is centred in the subfield of structure discovery and characterisation of complex adaptive systems that are described by adaptive networks and high dimensional relations. Although centred on structure finding, this work also embraces some other aspects of complex systems studies, namely coordination processes through stigmergy phenomena, showing how these can help understand and characterise adaptive behaviour in complex networks. Attention is focused is given on the area of hypernetworks and Q-analysis as the tools to describe the structure of the high dimensional relations existing in these adaptive networks. The study of this thesis was carried out through the illustration of real world case studies. The object of study are the news stories published by online newspapers, forming a corpus of written documental data whose structure this thesis aims to understand. From this structure better algorithms are developed that retrieve, digest and organise this information.

In this introduction the field of research is presented in section 1.1. In section 1.2 the object of this study is explained and it is followed by the objectives of the thesis in section 1.3. The scientific contributions are given in section 1.4, and finally the outline of the full

document is presented in section 1.5.

1.1 Field of research

Network science is an emerging field in complex systems. It studies the complex networks that arise naturally in many fields: engineering, biology, cognitive sciences, linguistics and social networks. The advent of the the world wide web revolutionised the way information is presented and how people access it. This created an opportunity for new complex networks to emerge, with new properties that derive from their intrinsic characteristics. Today's network science deals with high dimensional complex systems. Recently published papers mention "adaptive networks", "networks of networks", "multiplexes", and "hypernetworks". Now is the time when network science departs from simple binary models of relations into high dimensional descriptions of the interrelations between parts of the studied complex systems (Díaz-Guilera, 2013).

Historically, complex networks have been modelled either as random graphs (Solomonoff and Rapoport, 1951; Erdös and Rényi, 1959, 1960), small-world graphs (Travers and Milgram, 1969; Watts and Strogatz, 1998) or scale-free graphs (Barabási and Albert, 1999) following the findings on properties of those classes of graphs. In recent work, it has been argued that social systems (Hamill and Gilbert, 2008) tend to be poorly modelled by these network models and the field of network science is expanding to consider multi-level networks (Criado et al., 2010, 2012). Networks of networks have become an important research topic because many real social systems are better described by high dimensional relations (Atkin, 1974, 1981; Johnson, 2006b).

Traditionally, network science was concerned with understanding the structure of static networks, finding scale invariances in their topologies and fundamental laws for their generation. However, as the latest research shows, networks cannot really be studied synchronically as they are not static topologies of nodes and connections in which some process occurs. Networks coevolve in a coupled mechanism between their dynamics and their topology, both affecting each other over time, in what Atkin described as a co-evolution between backcloth and traffic. One does not exist without the other and each affects the other's behaviour (Atkin, 1974). This feedback mechanism leads to a new class of networks, called adaptive networks (Gross and Blasius, 2008). They still have the same characteristics as other networks, but the time-scale of the evolution of the topology

is similar to the time-scale of the process that occurs in the network. This makes those networks hard to study in a synchronic view and a diachronic approach is needed.

The existence of these dynamics leads to the necessity of understanding the feedback mechanisms and interconnections of the different levels of the complex adaptive systems. Such coordination between subsystems leads to the study of stigmergy phenomena, as it is clear that the mechanisms, which occur in adaptive networks, are not driven by external forces, but are self-sustained, and evolve dramatically over time. This form of self-organisation of adaptive networks is a characteristic found in many real problems, and many times it is mixed with noise that masks the real structures underlying (and driving) the system. This phenomenon was studied as it occurred in the context of ant colony optimisation algorithms (Dorigo and Gambardella, 1996) so as to understand the importance of this concept in adaptive networks and as a way to identify how the connectivity properties of the underlying structure can be used to navigate the high dimensional texts in a useful and coherent manner.

The application of graph theory is an interdisciplinary science, with results coming from many fields. While most of the results produced by science have been statistical in nature, graph theory as related to hypernetworks has a strong descriptive power for the relations existing in these complex systems. These adaptive networks were also studied as related to Q-analysis and hypernetworks. This approach is considered the descriptive language of structures. The questions studied in network science are essentially problems of relations between sets, and how these sets and relations evolve. For this reason the language of structure is very powerful, albeit different from a traditional statistical approach (Atkin et al., 1968; Atkin, 1972; Beaumont and Gatrell, 1982; Johnson, 1983).

This research tackles several topics currently important in complexity. The main problem of the thesis resides in the characterisation of the structure of online news. For this it explores many ideas in network science, where it focuses on complex adaptive networks. It also embraces other topics like stigmergy, self-organisation, information theory, community detection, and Q-analysis. Each has its part in the results presented here and all interweave a framework that allows for a better understanding of a modern adaptive system such as the online news publication.

1.2 Object of study

The object of this work is the analysis of written documental corpora published in online distribution channels by newspapers. The amount of electronic information presently available through electronic publications, e-books, emails, news, or blogs, is growing in contrast to the printed counterparts. This can be observed in the following figures.

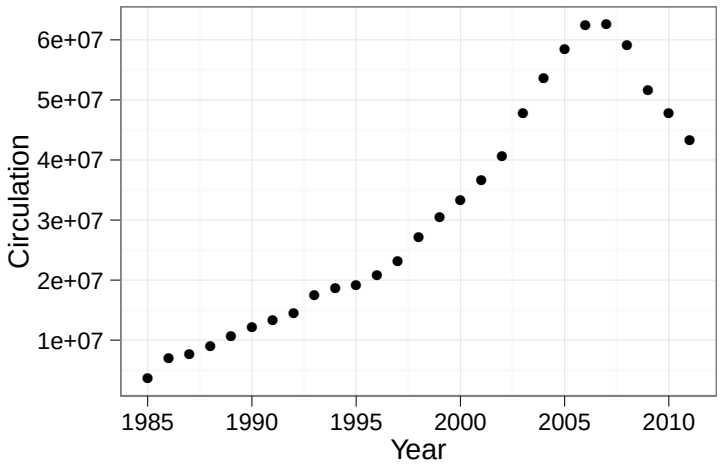


Figure 1.1: Portuguese journal+magazine printed circulation

Figure 1.1 illustrates the Portuguese Journal and Magazine Circulation numbers[1]. It is clear that after constant growth in publication numbers, after 2007 the circulation numbers have declined.

The same phenomena is observed in figure 1.2 for the aggregated circulation of the United Kingdom top journals[2]. In this particular case the circulation maximum was observed around 2002 and has since decreased.

On the other hand, Internet traffic has grown exponentially in the past 21 years, as shown in figure 1.3. This means more content is being published online then ever and naturally newspapers are adapting to this new medium.

This change of publication channels (from print to online) can be clearly observed in a 2008 study by the Pew Research Center for the People & the Press, that showed that the Internet overtook newspapers as the preferred news outlet by American readers (Kohut and Remez, 2008)(Figure 1.4).

[1]Data from the Portuguese Circulations Bureau - http://www.apct.pt/
[2]Data from the Audit Bureau of Circulations - http://www.abc.org.uk/

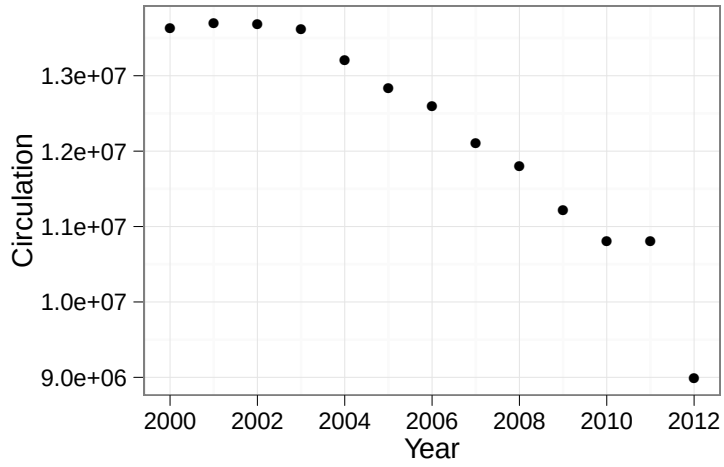


Figure 1.2: Aggregated UK 13 top journals daily printed circulation

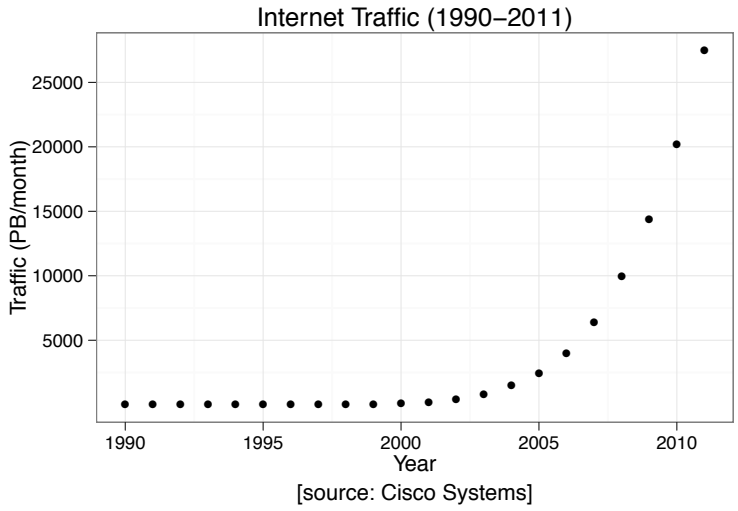


Figure 1.3: Internet Traffic in 1990-2011 (in PetaBytes/month)

Taking into consideration the data presented it is clear that this trend favouring online publication of news will continue and that it will be driven by the ubiquitous presence of connected devices and changes in the population's reading habits.

This change of paradigm for the publication and consumption of information is prevalent across the different domains of publication and communication outlets. The hypothesis defended in this work is that adaptive networks theory is an adequate framework for the understanding of online news systems, allowing a look into this system in an abstract way, free of the idiosyncrasies of each publication, language, political orientation,

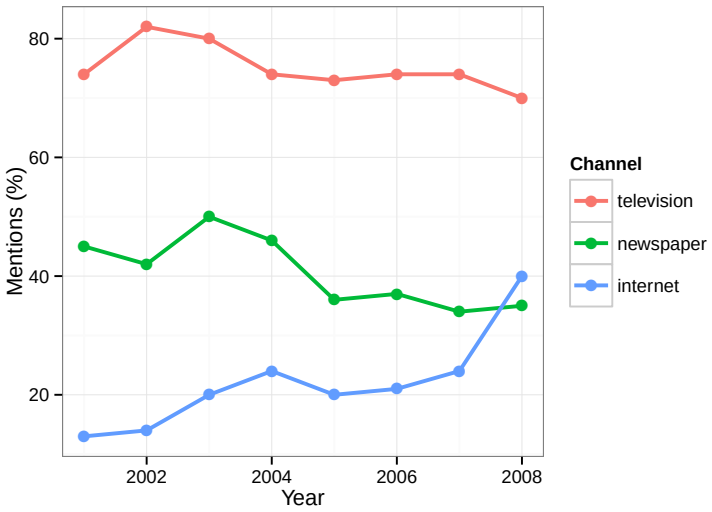


Figure 1.4: Internet overtakes newspapers as news outlet (Kohut and Remez, 2008)

or authorship. By transforming the corpora into abstract structures, based on algebraic topology, there is the advantage that all the techniques of data manipulation which are the best practices in the field can be used while keeping a proximity to the original data sources that a pure statistical manipulation would not allow.

After 20 years of technological advances, the Internet has become a favoured distribution channel of news to the detriment of the paper based publications. This new medium does not present the same characteristics as traditional print: from gathering, writing, publication and consumption of the news, all these processes evolved and naturally these changes must be reflected in the way news is published and presented in the new online channel. My objective is to present a contribution to the studies of these news systems, trying to show the features of these news systems.

The corpus of this research is made of news stories published for a time span of approximately 3 years, coming from a variety of online journals, written in different languages, and in different geographic regions of the planet. These were collected under the auspices of **The Observatorium**[3] laboratory. For the purpose of this work two newspapers were selected for detailed analysis. Through the analysis of these data, one may find the commonalities between different publications, understand how the news are treated by different media, and ultimately reveal the structure of the news published online. This interest leads to the formulation of some questions about this corpus: does news published

[3]http://www.theobservatorium.com - The Observatorium is a project running at ISCTE-IUL with the support of the Master and Doctoral Programs in Complexity Sciences.

in different newspapers have identifiable structural properties? Is it possible to detect patterns in these newspapers? Is it possible to detect novelty, topics, or trends from the data? These are some of the questions that motivated this research and the choice of this corpus for study.

Aiming to understand the documents produced by newspapers, with the minimum of human intervention, the original work of this thesis focuses on techniques that do not require *a priori* knowledge of the data. It focuses on the exploitation of the existing structural properties of the documents allowing the documents to be clustered, organised and digested in novel ways. This is achieved by devising methods and algorithms based on the connectivity of simplicies (in Q-analysis), information theory, bio-inspired methods, and adaptive networks.

1.3 Main objective of the thesis

The thesis objective is to clarify the structure of news published in online newspapers. To address this main objective the language and concepts of complexity science are used. For this, a set of hybrid approaches and multi-level analysis of the system are used, mainly Q-analysis (Atkin et al., 1968, 1971; Atkin, 1972; Johnson, 1976, 1978b), information theory (Meilă, 2007; Rodrigues and Louçã, 2009; Rodrigues, 2010), adaptive network theory (Gross and Blasius, 2008; Caldarelli and Garlaschelli, 2009; Gross and Sayama, 2009; Gross, 2009), and bio-inspired ant colony optimisation techniques (Dorigo and Gambardella, 1996; Dorigo and Stützle, 2009; Jafar and Sivakumar, 2010; Rodrigues et al., 2011; Ramos et al., 2013). The dynamic nature of the news based corpora make the task of pre-defining categories very difficult. Also, it presents the difficulty and subjectivity of labelling the documents, to produce a training set in this context and because of this, unsupervised approaches are used to tackle the problem. Clustering is unsupervised and automatic, but still subject to some challenges:

- the number of documents to cluster is usually high;
- the number of parameters is usually high;
- the number of clusters is difficult to establish *a priori*;
- the documents might belong to more than one cluster (overlapping);
- the structure of each cluster is arbitrary and can be significantly different from other clusters in the partitioning;

- the analysis must be executable in real-time or quasi real-time, due to the growing amount of documents produced presently.

It is my belief that these challenges can be met by the development of hybrid methods that combine several of the techniques highlighted before. By using the underlying idea that news is connected in some natural structure, I show how to take advantage of this natural connectivity and, by combining it with other techniques, produce hybrid methods that give insights into the problem of finding the structure of online published news.

1.4 Scientific contribution of this thesis

The outcome of this thesis, the algorithms and methods described, can be used in the detection of novelties in scientific publications, in the summarisation of documents for market studies, brand following in social networks, sentiment analysis, or voting prediction in elections. The corpora of documents can be easily collected from documents published online, being newspapers as in this thesis, or other kinds of documents that fit the study that the task demands. Any system that can be mapped into relations among different sets is subject to being treated by the methods and techniques presented in this thesis as they are context free and can easily be generalised to a variety of real world problems.

The work presented in this thesis contributes to the advance of knowledge in science in several aspects:

- In the field of adaptive networks it presents a novel way of analysing the structure of noisy co-evolving networks through the use of Q-analysis and community detection methods (Rodrigues, 2013). This combination filters noise from news and allows the extraction of the core relevant topical clusters from the corpus of collected news.
- A new algorithm for the resolution of combinatorial optimisation problems, that can be mapped to the well known travelling salesman problem and its variants (Ramos et al., 2011; Rodrigues et al., 2011; Ramos et al., 2013). This algorithm is an extension to Ant Colony Optimisation algorithms and is bio-inspired by the finding that some species of ants use a negative pheromone to mark some paths with non-entry signals. The benchmark results show that this approach is superior to the traditional Ant Colony System algorithm presented by Dorigo and Gambardella (1996). This algorithm was applied in the construction of Hamiltonian paths on the news corpus using their eccentricity as a measure of distance. This hybrid approach

reveals the importance of using the high dimensional connectivity of the systems to improve the results. The dual pheromone mechanism also opens doors for novel bio-inspired algorithms where second order reasoning might improve the quality of the analysis.

- It adds to the understanding of existing structure in online newspapers through the analysis of content published in websites (Rodrigues and Louçã, 2009; Rodrigues, 2010, 2013). Through the joining of Q-analysis with clustering techniques based in modularity optimisation, coherent news topics from the global ensemble of news published daily can be identified.
- As a necessity of the study, software called Theseus was developed that provides an easy framework for the processes of gathering, archiving, and analysing online documents. This software acts as a library that researchers can integrate easily when developing new algorithms, either by changing the ones implemented, or by adding new functionalities. The modular character of Theseus guarantees to new researchers quick development of new functionalities and greater experimental time. The software Theseus is presented in appendix A.

Table 1.1: Hybrid approaches developed during this thesis

Aspect	Chap. 8	Chap. 9	Chap. 10	Chap. 11
Variation of Information	✓			
Adaptive Network	✓			
Bags of Words		✓		
k-means Clustering		✓		
Modularity Clustering			✓	
Q-analysis			✓	✓
Bio Inspired				✓

Table 1.1 highlights the hybrid connectivity approaches developed in this work, showing in which chapters of the document they can be found. In Chapter 8, the detection of topics was accomplished by connecting an information based measure (variation of information) with an adaptive network where each node is being added to the evolving network of news. In Chapter 9 a traditional bag of words model produced feature vectors that were then clustered via a traditional k-means method to produce timelines of the news. The connectivity of the systems described by Q-analysis was used as a filtering mechanism for the generation of a network, then clustered via modularity optimisation in Chapter 10. Finally a new biologically inspired algorithm based on ant colony systems was coupled to the Q-analysis description of the news in Chapter 11.

Table 1.2: Contributions of this thesis according to problem

Problem	Scientific Contribution	Software Contribution (Theseus)
Tracking Trends in News	Use of Variation of Information with Adaptive Networks to find clusters	Visualisation of dynamics of networks while tracking trends
Finding Modules in News	Use of Modularity optimisation on Q-analysis descriptions of bipartite graphs	Theseus, software to compute Q-analysis descriptions. Visualisation software for clustered timelines of news
Filtering 'Noise News'	Use of Q-analysis to filter low connected nodes in bipartite graphs	Visualisation software for the structure of the Q-analysis description of the bipartite graph
Finding Hamiltonian Paths	Using bio-inspired algorithms combined with eccentricity from Q-analysis description of bipartite graph. New Ant Colony algorithm using negative pheromone as a second order no-entry marker.	New ant colony algorithm. Visualisation of Hamiltonian paths of news for web navigation through them.
Retrieving Published News		Develop a crawler Theseus / Processor for online document retrieval

Table 1.2 summarises the contributions of this work both in terms of the problems that were solved in the analysis of the structure of news published online and in terms of the scientific and/or software contributions produced with this research. Besides the software contributions, the scientific contributions show how the connectivity of the underlying structure is explored to solve the main challenges posed by this problem.

1.5 Thesis outline

This document was structured according to the thesis and dissertation plan defined by the *International MSc and PhD Programs in Complexity Sciences*[4] and follows the presentation norms for a PhD thesis of ISCTE-IUL.

After this introductory chapter, the thesis is divided into two parts followed by the conclusion in Chapter 12 for a total of 12 chapters. After the conclusion, there is a third part with several appendices relevant to this work.

Part I (Chapters 2 through 7) presents a review of the related work relevant to this work. Chapter 2 presents a brief literature review on how this problem of topic finding in textual analysis is traditionally handled. It is followed by a review of adaptive networks in Chapter 3, the main works of Q-analysis are shown in Chapter 4, the community detection methods are presented in Chapter 5, and ant colony optimisation techniques are reviewed in Chapter 6. Ending this first part of the document, Chapter 7 summarises the research opportunities that are explored in this thesis.

Part I aims to give an overview of the most relevant work in the field of this thesis. However, while the main publication of these fields are listed in Part I, Part II includes specific bibliographic references when appropriate.

Part II of this thesis consists of Chapters 8 through 11. It presents the proposed hybrid methods and their applications to the case studies. It introduces the work done on the analysis of the structure of online news in selected case studies and with different approaches to reveal the structure of the news. It focuses on the description of the problem and describes the solutions developed for gathering the online newspapers. It also describes the dataset and the case studies chosen for analysis in subsequent chapters.

The first case study is the analysis of the news from the newspaper **Público** in Chapter 8, where a new method combining variation of information and an adaptive network is used to find the trending news in an automated way.

The second case study explores the news published by **The Guardian**. In this case several approaches were used: in Chapter 9 a timeline construction of the news connected to clustering by k-means; in Chapter 10 a method to filter noise from news and cluster them. In this context noise is defined as sparsely connected news that are unrelated to the main topics being published by the newspaper and that are connected to the structure through

[4] Website: http://www.complexsystemsstudies.eu/

very general tags. In this method the application of Q-analysis filters the noisy news and then the resulting subgraph is clustered by a modularity optimisation algorithm. Finally, a novel algorithm for finding Hamiltonian paths in the travelling salesman problem (TSP) is presented in Chapter 11. This algorithm is based on ant colony optimisation meta-heuristic and the eccentricity from Q-analysis description is used to produce an hybrid algorithm that finds a walk through related news published by **The Guardian**. The algorithm solves the TSP problem through Ant Colony Systems and exposes how the introduction of a negative, 'no-entry' signal, pheromone improves the performance and convergence of the system. In this chapter, the role that the second order stigmergy variable (negative pheromone) plays as the driving force in finding optimal solutions is also discussed.

Finally, Chapter 12 concludes this work with the review of the main results found in the case studies. The limitations of this methodology are discussed, new opportunities for further research are presented and the importance of the hybrid methodologies in high dimensional complex datasets is pointed out.

Part III contains the appendices of this thesis, namely the description of the software developed for processing and visualisation of the results presented in this work in appendix A. This part also includes two appendices (B and C) that are listings of results from the algorithms developed in Part II.

Part I

Related Work

In this part, a review of the related work in the main fields of this thesis is presented. It starts in Chapter 2 with a brief overview of some of the approaches that have been used to deal with the analysis of textual corpora, namely in the fields of term extraction and automated text clustering. These extraction techniques aim to understand documents produced in ways manageable by machine algorithms for further analysis. An overview of Adaptive Networks is presented in Chapter 3 as they are used for the extraction of news clusters in the case study of *Público*. In Chapter 4 a review of the main aspects of the mathematical language of structures and relations, Q-analysis, is given, as this will be extensively used in the analysis of the case study of *The Guardian*. In Chapter 5 a review of the main community detection methods and strategies is presented. Community detection methods are clustering methods on graphs and are of interest in the case study of *The Guardian* to aggregate news into coherent clusters. This is followed by Chapter 6 where ant colony optimisation systems are reviewed. This chapter highlights some of the most important novelties in the field of bio-inspired optimisation techniques. The importance of these is clear as a new ant based algorithm was developed in this thesis, that is applied to finding the structure of news in subsequent chapters. This part finishes in Chapter 7 which aligns these fields with the thesis perspective and indicates research opportunities by combining these different fields to produce new algorithms that solve the problem of identifying structure in news published online.

Chapter 2

Analysis of news

In recent years there has been a surge of interest in term extraction and automated text categorisation (Yang and Liu, 1999; Pantel and Lin, 2001; Cardoso-Cachopo and Oliveira, 2003), mainly because of the increasing amount of information that is being produced online. Examples include extracting and classifying biological text (Lee et al., 2006), the categorising of online news (Jo et al., 2000; Weninger and Hsu, 2008) and personalised recommendation systems (Zhang et al., 2010). With online information's rapid growth also has come the development of automated and agile methods to process it. The literature provides many examples of term extraction methods that can be categorised roughly into two fields, according to their different perspectives. One approaches the task from a linguistic, terminology and natural language processing perspectives (Hatzivassiloglou et al., 2000), and the other uses mainly tools from the statistical and information retrieval fields (Nigam et al., 1999; Pantel and Lin, 2001).

One of the great challenges in the process of term extraction is related to the peculiarities of the language being processed. Traditional evaluation relied on assessments made by humans about the characteristics of the extracted terms. This evaluation method can be difficult to apply to large datasets. By combining several automated strategies the aim is to reduce human intervention to the bare minimum.

The high volume of data collected in these systems demands mechanisms to extract relevant information in ways that are quick, relevant, and automatic. Therefore it was necessary to develop tools for automatic processing of these documents. Two kinds of document organisation are used presently, and they affect the kind of methods used accordingly. The first kind is **categorisation of documents**, which aims to give a new document a label according to a previously defined ontology. The second kind is **clustering of documents** which aims to group documents together without any pre-constrains. Typically, these two

classes of methods correspond to two types of machine learning algorithms: categorisation is usually done with supervised learning algorithms and clustering is usually done with non-supervised algorithms.

In the **categorisation of documents** field, a series of pre-defined labels are attributed to documents. This attribution is based on some estimation of similarity between the the attributions made and the attributions of a training set. This process requires human intervention in the definition of the labels, construction of the training corpus, and the definition of the rules that attribute certain labels to certain documents. Sebastiani (2002) discusses several machine learning techniques for automatic classification of text. The author concludes that these machine learning techniques usually lead to a time saving, when compared to traditional methods where classification is accomplished by knowledge engineering. A group of specialists has to define the rules for the document classification in such cases. Several algorithms have been developed for the classification of documents that include: *k*-nearest neighbour (*k*-NN) (Fix and Hodges, 1951), support vector machines (SVM) (Vapnik, 1982; Cortes and Vapnik, 1995; Joachims, 1998), Neural Networks (NNet) (Wiener et al., 1995; Ng et al., 1997; Ruiz and Srinivasan, 2002), linear least squares fit (LLSF) (Fuhr et al., 1991; Yang and Chute, 1994), and naive Bayes (BNB) (Lewis and Ringuette, 1994; Tzeras and Hartmann, 1996). A comprehensive comparison of these **classification** techniques can be found in the works of Yang and Liu (1999) and Sebastiani (2002).

In recent years interest has surged in the area of topic spotting, sometimes also called trend tracking. Researchers have applied an increasing number of learning approaches, including regression models, nearest neighbour classification, Bayesian probabilistic approaches, decision trees, inductive rule learning, neural networks, on-line learning and Support Vector Machines (Joachims, 1998; Yang and Liu, 1999; Cardoso-Cachopo and Oliveira, 2003; Hamamoto et al., 2005; Miao and Qiu, 2009; Solé et al., 2010). Most of these methods are supervised and require a training set where documents previously classified by humans are used as input to make the system learn each category's particular features. This approach poses two main problems: the need for a language-dependent analysis and classification by specialists and the difficulty in finding new categories. A new text is either part of one of the existing categories or not a part of any of them at all.

By contrast to **classification** methods, the unsupervised machine learning methods are mainly aimed at **clustering of documents**. In these tasks, usually pre-labeling of

documents is unnecessary, neither are sets of training documents. The objective of these clustering techniques is to group documents into subgroups in such a way that documents have high similarity between them within a subgroup and low similarity to other subgroups. The clustering techniques have been applied to diverse areas: information retrieval, where they were used to improve precision and recall; hierarchical classification of documents; document navigation for search engine optimisation (Brin and Page, 1998; Dumais and Chen, 2000; Beil et al., 2002; Hamamoto et al., 2005). Clustering algorithms can be organised into two main subgroups: hierarchical clustering and partitional clustering. In the former case one obtains subclusters inside subclusters, forming a dendrogram, by divisive or agglomerative methods. In the latter case, the process aims at clustering the corpus into subsets of non-overlaping elements without building a dendrogram.

The study of the clustering problem precedes its applicability to the text domain and was mainly focused initially on quantitative data in which the attributes were numeric. Text clustering finds applications in several tasks. Namely document organisation and browsing where hierarchical organisation of documents into coherent categories is useful for systematic browsing of the document collection; corpus summarisation where clustering can be applied to perform cluster digests of sentences or of words; and naturally clustering can help the previously mentioned task of document classification by the use of word clusters as input features for classification.

Several characteristics make text clustering a special case of clustering techniques that require special attention when compared with the traditional clustering methods (Aggarwal and Zhai, 2012).

- The dimensionality of the representation of the text documents is very large, but the data is sparse. The underlying lexicon from which the document is constructed can be very large ($\sim 100,000$ words), but on the other hand a typical news document might be only a few hundred words. This leads to large sparse matrices representing the data.
- The words are usually highly correlated with each other. This means that the concepts present in the text are usually smaller than the feature space. This requires the careful design of algorithms that take into account the specificity of the problem.
- It is usually necessary to normalise the documents by their size as comparing them directly is usually not possible due to the differences of size.

One traditional algorithm that has been proposed by MacQueen (1967) over the initial

idea of Steinhaus (1957) is the *k*-means clustering algorithm. This is popular method for clustering a number of observations into *k* clusters from the vector representations of those data cells. This algorithm lead to the construction of clusters where each observation is closest to that cluster mean and is therefore inside a Voronoi cell.

Dharanipragada et al. (1999) presented two online algorithms for story segmentation and topic detection based on decision trees, that combine machine learning with statistical natural language processing and information retrieval techniques. The topic detection algorithm is an incremental clustering algorithm that is implemented after a decision tree based probabilistic model. This work acts on streams of text produced mainly for broadcast speech recognition. In the area of text segmentation, maximum entropy based methods were proposed by Beeferman et al. (1999) where trigrams were used to segment news published both on the Wall Street Journal and on transcripts of news broadcasts. These two works are interested in finding the location of where a topic changes inside a stream of text, instead of finding clusters in a corpus of many documents.

Topic detection and tracking was mainly developed in the early XXI century as a funded program to research automated methods of finding event based organisation of broadcast news (Allan, 2002).

Another approach to this problem of topic detection and tracking has been proposed by Lavrenko et al. (2002) where relevance models were applied. The authors describe how relevance models were applied to datasets originating in the American press industry (two newswire sources, Associated Press and The New York Times, two radio sources, Voice of America and Public Radio International, and two television sources, CNN and ABC) in a corpus with more than 40,000 news stories.

McKeown et al. (2002) presented an online summarisation system that combines different techniques from topic detection, text clustering and summarisation to crawl websites, filter out news from non-news, group the news into stories of the same event and create summaries of those events for publication to the readers of the system.

Allen et al. (2007) presented a framework for the analysis of historical newspaper articles, by taking into account the necessities of optical character recognition (OCR). In their work they focus mainly on automated feature extraction from the OCRd text that would facilitate automatic indexing.

Feng and Allan (2009) presented a system that automatically analyses and presents news to the reader in a synthetic way. For this they explore the notion of *incident threading*.

All text that describes an event is collected into a news incident. Incidents are then organised in a network with dependencies of predefined types. The notion of incident relies on a journalistic description of events as it defines the Who, When, Where, and What of each event. Then, events are coupled with a contextual relation through a predefined vocabulary of the possible relation types, to build an incidence network of the events.

Zhou et al. (2009) proposed a method of graph clustering based on structural/attribute similarities that integrates structural and attribute-based clustering by adding attribute vertices to the network in addition to the original structural vertices. The same augmentation process is done to edges and the resulting augmented graph is a semi-bipartite graph. Then the authors proceed to perform a neighbourhood random walk in order to determine the closeness of the vertices. This closeness is then used to perform the clustering of the documents. This algorithm's main challenge is then to balance the weight of the relative structural and attribute components in the clustering process. The problem of finding clusters in networked data combining link and content analysis was also discussed by Yang et al. (2009). In their work the text content of a social network is attached to the nodes and a conditional model for link analysis is proposed. Then a second discriminative model helps reduce the impact of irrelevant content attributes and these two models are unified via the community membership.

Kim and Oh (2011) presented a new framework based on probabilistic topic modelling for uncovering trends and structure of important issues found in the news archives published online. Central to their approach is the notion of a topic chain, a temporal organisation of similar topics. The authors discuss how these topic chains are constructed and their meaning in the context of 9 months of Korean Web news.

Recently Cheng et al. (2012) performed a fine-grained topic detection on news search results from online search engines. Their approach to perform the topic detection on news search results is accomplished via an agglomerative clustering algorithm and a simulated annealing for optimisation of the process. Also, Liu and Chang (2013) prosed recently topic detection and tracking based on a new word measurement scheme called TF-Density that is modified from traditional TF-IWD and TF-IDF models.

Al-Kabi et al. (2012) recently proposed a novelty detection system for news published in the arab language. The system is able to cope with the automated processing required by the high volume of news published online in Arabic news websites and is able to identify the topics being covered by those news.

As shown here the problem of text analysis has been a popular research subject over the past 20 years. The massive production of text based documents and news that is accessible with the explosion of the Internet led to new ways of analysis to find structure and meaning in these datasets. In recent years, the main trend has been centred on research that copes with the large amount of text being created by dynamic applications such as social networks or online chat applications. Also in the future it is expected that many more cases of text analysis will move from pure text clustering to mixed applications where photos, links, audio, video and other elements will be of upmost importance. It will then be necessary to adapt existing text algorithms to deal with these heterogeneous scenarios.

The references presented in this review show relevant works both in the problem of identifying topics, or related events, and also the application of these techniques in building solutions that can benefit the final user. A comprehensive review of text clustering algorithms is available from Aggarwal and Zhai (2012).

The consumer of news in the end just wants to be better informed. Automated techniques that fetch, digest, and summarise the news are of importance now when the greatest rate of information production surpasses human capacity to manually search and read all available information.

Chapter 3

Adaptive networks

3.1 Network science introduction

This work builds on top of several fields of complexity theory, the most important of which is the field of networks. The research on networks has a long history and probably its most famous application being the solving of the Koningsberg bridges problem by Leonhard Euler (1741) that usually marks the date of the beginning of the studies in graph theory.

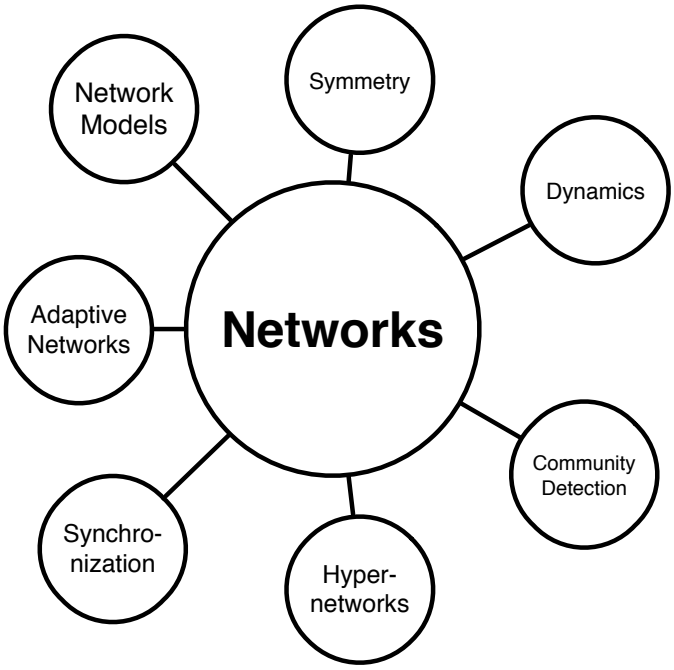


Figure 3.1: Main areas of network science research

In network science, complex networks can be defined as a collection of vertices

connected by edges representing various kinds of interactions (mainly non-linear) among the vertices. Today we are presented with many important large scale systems that can be represented by a complex network of interacting entities. Networks like the Internet and the power grid are becoming ever larger, encompassing millions if not billions of vertices. They exhibit complex and often dynamic patterns of edges between the nodes. These systems are very often organised in multi-level structures of interacting complex networks. Social networks, for example, are built upon information networks which are built upon communications networks which in turn are built on physical networks. This makes network science a broad area where many researchers from different disciplines converge, many times with divergent languages and definitions of the same notions (for example vertices are called nodes or sites or actors in different disciplines and contexts). A literature review of this area is daunting as the task would require an encyclopedic approach and could not be accomplished in the limited time and context of this work.

The field of network science is mapped in figures 3.1 and 3.2 (detailed map). In recent years there has been great interest in the field of adaptive networks and the number of scientific publications in this field has increased substantially. This opened the field to the need for some kind of classification system of all scientific production in order to better understand it.

As seen in the map of figure 3.1 the field of networks has expanded and encompasses several topics from other areas. It is important then to group this ensemble into topical areas. In figure 3.1 these are shown grouped for clear reading. In figure 3.2 the classification presented in figure is expanded for a detailed view of the area of network science.

As this field is vast, this section does not present an extensive review of network science. A brief introduction to the field is given, followed by the important aspects of adaptive network theory that are relevant to this thesis. Interested readers are advised to consult Albert and Barabási (2002); Dorogovtsev and Mendes (2003); Diestel (2005); Caldarelli (2007); Dorogovtsev (2010), and references therein, for a deeper review of network and graph theory. Dorogovtsev and Mendes (2003, pp. 237-240, Appendix H) is recommended for a more comprehensive guide to the network literature.

Some basic notions of network theory are introduced in this section to allow a better understanding of other concepts later in this work. It starts with the definition of networks and will show how several classes of networks have been studied. These are presented following an historical perspective and will give an introduction to the field of networks.

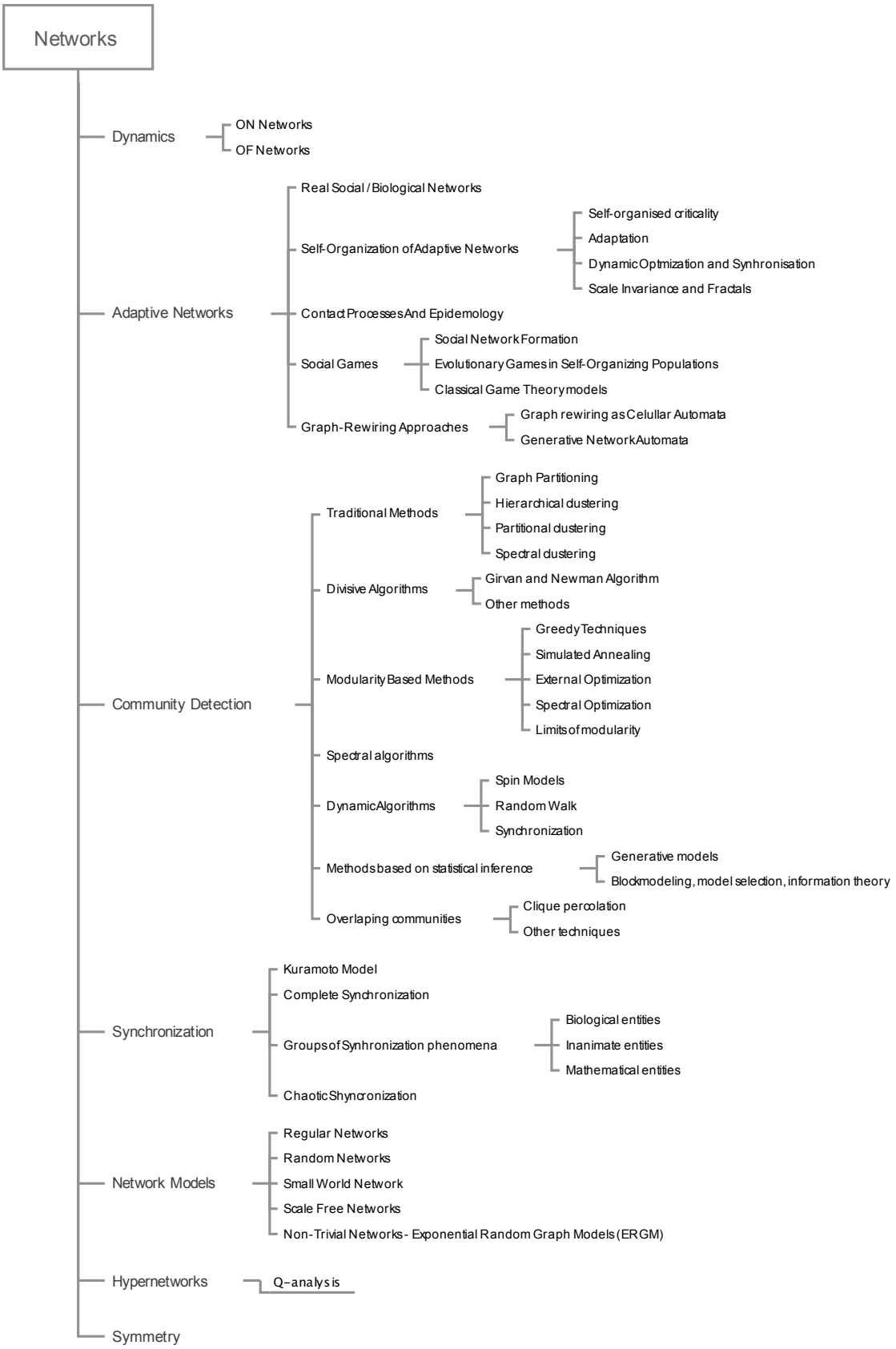


Figure 3.2: Map of the scientific areas of interest related to networks

3.2 A classical view of networks

$$G = \{Vertices, Edges\} \tag{3.1}$$

Formally, a network can be thought of as a pair of sets (eq. 3.1): a set of vertices (also called nodes), and a set of edges (also called links) connecting two elements of the vertices set. In its simplest form, a network is usually: *undirected* where the edge does not represent a directional relation between the nodes; *unweighted*, meaning that every edge between two vertices is topologically equivalent; has *no loops* (no edges connect a vertex to itself); and there are *no multiple edges* between any pair of vertices.

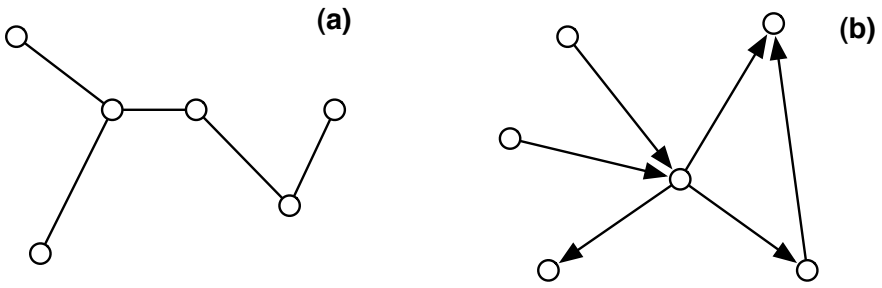


Figure 3.3: Sample networks: (a) Undirected graph were vertices are connected via undirected edges. (b) Directed graph.

The total number of connections that a vertex establishes is called its *degree* (see Fig. 3.3) and in the case of *directed* networks one can divide this notion into *in-degree* (the number of incident edges that a vertex has), and *out-degree* (the number of outgoing edges that a vertex has).

One way to completely represent a network is by the use of an *adjacency matrix*. It indicates which of the vertices are connected. This is a square matrix $N \times N$ where N is the total number of vertices. The entries of the matrix $a_{i,j}$ are 1 if vertex i is connected to vertex j and 0 otherwise. In the case of *undirected* networks this matrix is symmetric and $a_{i,j} = a_{j,i}$.

3.2.1 Regular graphs

This is the simplest form of graph. In regular graphs the topology of the network is a lattice where topologically there is no distinction between the properties of the different vertices. All present the same degree and establish local connections to neighbouring vertices. One vertex is representative of the entire set of vertices and the properties of the network can be derived from it.

3.2.2 Random graphs

In the late 1940s, early 1950s, Rapoport dedicated his research to the fields of mathematics and biology. This was a time when these two areas were highly separated (Newman et al., 2006). He was very interested in understanding networks more by their aggregate properties rather than by their individual nodes or connections. He co-authored with Solomonoff a series of papers in which the systematic study of random graphs first appears (Solomonoff and Rapoport, 1951; Solomonoff, 1952). The model presented become known as the Gilbert model of a random graph where the notation $G_{N,p}$ indicates that this is a statistical ensemble of networks, G, with two fixed parameters: the number of vertices N in each ensemble member, and a given probability p that two vertices have an interconnecting edge.

Following on from Solomonoff and Rapoport (1951) and Solomonoff (1952), in the 1960s, Paul Erdös and Alfréd Rényi also worked on the study of random graphs. In the Erdös-Rényi model the statistical ensemble of random graphs is denoted by $G_{N,L}$ where N defines the number of vertices and L the number of edges in those graph realisations. Through their study of random graphs they showed that the evolution of the topology of the graph as the number of connections increases changes dramatically. They showed that the properties of these random graphs appear suddenly (e.g. the giant component appears when the mean degree $\langle q \rangle > 1$), going through a phase transition and not gradually, as one might expect at first (Erdös and Rényi, 1959, 1960, 1961).

The two models for random graphs are not equivalent as the Erdös-Rényi model allows for the existence of multiple connections and loops. In any case when considering graphs with large numbers of vertices $N \to \infty$ while keeping the mean degree $\langle k \rangle$ constant, the two models became equivalent and the models approach a Poisson distribution for the degree k of the network's vertices:

$$P(k) = e^{-\langle k \rangle} \langle k \rangle^k \frac{1}{k!}. \tag{3.2}$$

3.2.3 Small-world networks

According to Schnettler (2009, p. 166), the 'small world research is now commonly associate with a manuscript of Pool and Kochen (1958) that circulated around 1958 but was not published until 20 years latter'.

Many networks that are observable in nature are compact, they present very short average path lengths when compared to their size. This effect was studied by Milgram (1967); Travers and Milgram (1969) in what became known as the Milgram experiment or the *small world* experiment. Milgram wanted to learn about the probability that two randomly selected people would know each other. The idea of the experiment was that randomly selected people in cities in the USA would be given a package along with information on who to send the package to. These people were asked to forward the letter to the destination subject if they knew him or her on a first-name basis. Otherwise, they would send the package to someone they knew on first-name basis that they thought could deliver the package to the final recipient. The experiment was designed so that it was possible to measure the path lengths between initial and destination people and the conclusion showed an average path length of 6 making the researcher state that people in the United States are separated by about six people, on average. Although these numbers have later been revised, this gave rise to the expression "six degrees of separation" (a phrase never used by Milgram or his team). The take home message of his experiment is that real networks are indeed compact and therefore exhibit some characteristics that researchers tried to mimic in graph constructs.

The random networks have particular interesting aspects. One of which is that they do not have big diameters when compared to the number of nodes. They are compact. In fact the long range interaction assure that these networks have small average path lengths. Another feature is that these networks have very few triangles and therefore their clustering coefficient is very small. On the other hand, lattices usually have many triangles, but no long range connections and their average path length is large. By combining these two types of idea Watts and Strogatz (1998) developed a method to build networks that presented short average path lengths while still containing many triangles. Starting from a

lattice, one then proceeds to randomly rewire, according to a probability β, the edges in the lattice. This results in a series of shortcuts that will connect to different parts of the original lattice, in practice shortcutting the network, while still retaining many triangles in the network. These, more compact, networks where called by the authors *small-world networks*.

For a comprehensive review of small-world networks theory and history please read Schnettler (2009, and references therein).

3.2.4 Scale-free networks

The Erdós-Rényi model has been widely used in in the study of random graphs. However, several studies showed that this model does not capture the richness of structures encountered in real world networks and fails to reproduce many of these real world network properties. In systems composed of many elements one usually finds that the degree distribution of networks constructed from the relations observed in them, follow, for large values of degree k, a distribution function $P(k)$ that is a power law:

$$P(k) \propto k^{-\gamma} \tag{3.3}$$

This usually means that the structure presents few vertices (called 'hubs') with many edges and many vertices with few edges. Usually the parameter γ is in the range $2 < \gamma < 3$ but occasionally might be outside these bounds.

Examples of systems with this kind of structure abound in natural and social systems. The World Wide Web (WWW) is one such example. The first significant result finding on the distribution of connections in the WWW, was presented by Albert et al. (1999) where they mapped the connections on the top domain nd.edu, and showed that the network possess different power-law behaviours for the in-degree and out-degree of HTML hyperlinks between websites with characteristic exponents of $\gamma_i = 2.1$ and $\gamma_o = 2.45$. Later, Broder et al. (2000) analysed the structure of the entire WWW by crawling the links on the AltaVista[5] search engine to confirm the power-law behaviour, although the value of γ found in the empirical data for the out-degree diverged from the previous work by

[5]http://www.altavista.com/ - AltaVista was at the time considered to be one of the more complete search engines available. Today it is inactive redirecting results from the Yahoo search engine.

Albert *et al.*: $\gamma_i = 2.1$ and $\gamma_o = 2.7$.

A model for the construction of scale-free networks was presented by Barabási and Albert (1999); Albert et al. (2000, 2001) and follows a preferential attachment in what has become known as the Barabási-Albert model.

This model reproduces the time growth of many networks (including the Internet and WWW). The network is built in successive time-steps where vertices are added to the network and then connected to existing vertices according to a probability that is proportional to their degree at that moment. This leads to the construction of *rich-get-richer* kind of networks that mimic the topological growth of real networks.

For a comprehensive review of scale-free networks theory and example cases, please consult Caldarelli (2007, and references therein).

3.3 Theory of adaptive networks

While the studies in subsection 3.2 were interested in synchronic views of networks, which are communities, degree distributions, centralisation of actors in the networks, and network percolation. This section highlights the dynamical aspects of networks. A diachronic view is very important in the sense that one does not have a complete view of what a network process is if the analysis is restricted to a slice in the timeline of the network evolution. Also, the classification of adaptive networks can be made according to several contexts, or sub-themes. Figure 3.4 shows a classification of adaptive networks scientific publications by Gross (2009) were the diversity of research aspects (in the study of adaptive networks) is presented.

This classification is vast and aims at being a comprehensive classification of scientific publications aiming at a broader audience that will then navigate through this classification according to a subset of the classification tree.

This chapter takes into consideration two aspects of Adaptive Networks in terms of their dynamics (Gross and Sayama, 2009). First the dynamics that occurs on networks is discussed. This relates to information propagation over a defined structure, as for example in rumour spreading or epidemic contagion, where the nodes of the network see their properties changed over time. The second, discuses the dynamics of networks themselves. In this case the topological perspective of the network dynamics is of interest. How do the weights of different connections evolve? How are new connections formed and lost over

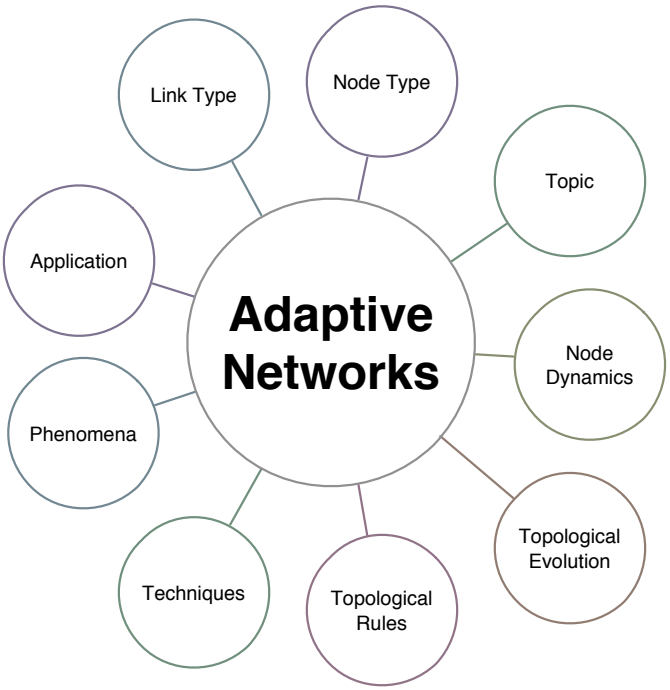


Figure 3.4: The classification of Adaptive Networks publications

time? These dynamics will help understand how some networks are more resilient over time or spread information more quickly than others. Adaptive Networks combine these two approaches.

3.4 Dynamics ON/OF networks

It is important to distinguish Adaptive Networks from Networks where some kind of dynamics is present in the form of dynamics ON networks or dynamics OF networks, but not both. Many studies focus only one aspect of the dynamics while the time-scale of the other aspect is considered irrelevant, or they are not considered interdependent.

The line of research that focuses on the *dynamics on networks*, considers that each node of the networks acts as a dynamical system by itself. The individual systems are then coupled to other dynamical systems according to the network topology where they are inserted. In these cases the topology of the network remains static while the dynamical processes inside the nodes evolve over time (Gross and Blasius, 2008). Examples of these systems include diffusion processes, percolation, epidemic propagation, contagion and rumour spreading, social influence, gossip, and voting models (Broadbent and Ham-

mersley, 1957; Hammersley, 1957; Albert et al., 2000; Pastor-Satorras and Vespignani, 2001),(Newman, 2003, and references therein),(Saavedra et al., 2007; Bjell et al., 2009).

In the case of *dynamics of networks* it is the topology that is regarded as a dynamical system. Usually topology evolves according to specific, often local, rules. Research in this process has revealed that different evolutionary rules give rise to different topologies with special properties (Gross and Blasius, 2008). These include for example the formation of 'small-world' networks or scale free networks (Watts and Strogatz, 1998)(Albert and Barabási, 2002). The dynamics of networks research is focused in studying topological changes, structural changes, network robustness, and network resilience (Albert et al., 2000; Dorogovtsev et al., 2002).

3.5 Defining adaptive networks

As shown previously, network analysis is usually carried out taking into consideration several aspects of the dynamics that happen on the network and that happen to the network topology. Studies of networks traditionally are centred in one of these aspects neglecting the other. Very few works explore both aspects simultaneously. When they do, they tend to explain how one aspect of the dynamics evolves by fixing the other. As an example some works freeze the topology, trying to explain how it affects the dynamic of a dynamical process on the network.

This kind of analysis revealed that different network topologies potentiate different properties at the level of the processes dynamics. For example it is well known that *scale-free* networks are typically resilient to random attacks and allow for the fast dissemination of information. But these cases are not examples of adaptive networks. Consider the case where one has a system made of two interlocked processes: 1) a dynamical process that occurs on the networks nodes and edges; 2) the process that continuously evolves the network topology. These two subsystems evolve in parallel simultaneously and happen in natural systems. What usually is also observable is that the time scale of these two subprocesses is sufficiently different that for all practical purposes one of the subprocesses is considered invariant on the time scale of the other. For example, the epidemic spreading of the flu virus on planetary scale through the air flight network is a case where the dynamical process on the network (virus spreading) occurs on a time scale much shorter than the time scale of the topological dynamics (the scheduling of flights is usually done

several months before the flight dates). These two processes also seem independent of each other. This system is not one of an adaptive network because the time scales of the two subprocesses are very different and the two subprocesses do not have any coupling mechanism (a kind of feedback).

In the other hand, if one considers the case of a system of pandemic transmission of a deadly virus through a network of social contacts, then it is a system where the temporal scales of the two subsystems are similar. The contagion can affect the survival of the people on the network and in this way alter its topology, or someone realising him or herself as carrier of the virus, might avoid establishing contacts (breaking edges) to avoid the spreading of the virus. In this situation it is clear that one can not speak of two separate subsystems. In reality the dynamic of the processes and the dynamics of the topology are interlocked and are interdependent, and it is best to consider this as single system.

The time scale of the process occurring on the network is comparable to the time scale of the topological changes and because of this none can be considered invariant to the other changes. Also, it is clear that both subprocesses are related and are interdependent. In the deadly epidemic example, one might argue that changes in the contact network occur in a similar time scale, even if the virus is absent. In any case, here it is considered that there is active action of the members of the network to avoid the virus by breaking connections. Also the death of people effectively removes nodes, and their connections, from this network.

Following the previous paragraphs, the Adaptive Network notion is now clarified. A system can be defined as an Adaptive Network, when both the subprocess of topological evolution and dynamical processes are evolving in such timescales that the study of one can't be done without considering the other. In the case of Adaptive Networks one usually observes that these subprocesses are:

- highly interconnected and interdependent
- one of the processes is a dynamical process that runs on top of the underlying topological structure
- one of the processes promotes the topological evolution of the network
- the time scales of the two subsystems are similar and can not be considered invariant to the changes of the other subprocesses
- there is a coupling between the events in one subprocess and the other that act as a feedback mechanism. Changes in the topology affect the dynamical process

evolution and this evolution change the way the topological process evolves. Both subprocesses co-evolve.

3.6 Some characteristic themes of adaptive networks

Gross and Blasius (2008) reviewed recent developments combining topological evolution of the networks with the dynamics in the networks nodes. They found that all these studies are characterised by common themes. Despite their diverse range they are able to synthesise them into four hallmarks of adaptive behaviour:

- "Self-organisation towards critical behaviour. Adaptive networks are capable of self organising towards dynamically critical states, such as phase transitions. This frequently goes together with power-law distributions. This mechanism is highly robust."
- "Spontaneous 'division of labour'. In adaptive networks, classes of topologically and functionally distinct nodes can arise from an initially homogenous population. In certain modes a 'de-mixing' of these classes is observed, so that nodes that are in a give class generally remain in that class."
- "Formation of complex topologies. Even very basic models of adaptive networks that are based on very simple local rules can give rise to complex global topologies."
- "Complex system-level dynamics. Since information can be stored and read from the topology, the dynamics of adaptive networks involves local as well as topological degrees of freedom. Therefore the dynamics of adaptive networks can be more complex than that of similar non-adaptive models." (Gross and Blasius, 2008)

The authors also make a first attempt at an inventory of dynamics of adaptive networks where they characterise them according to different paradigms (Gross and Blasius, 2008):

- "**Activity disconnects**. Rule: Frozen nodes gain links, active nodes lose links. Outcome: Self-organisation towards percolation transition, active nodes scale as power law."
- "**Like-and-like**. Rule: Connections between nodes in similar states are strengthened. Outcome: Heterogeneous topologies, possibly scale free networks, emergence of topologically distinct classes of nodes."
- "**Differences attract**. Rule: Connections between nodes in different states are strengthened. Outcome: Homogenous topologies, power-law distributed link weights."

In the book *Adaptive Networks* (Gross and Sayama, 2009), the authors divided the field of Adaptive Networks into several categories of research. The categorisation followed by the authors of this book gives the categories outlined next:

- Application Studies - Mainly papers that deal with social group dynamics in dynamic networks, community detection, and evolving communities.
- Time-dependent Complex Networks
- Biological Adaptive Networks
- Self-organization of Adaptive networks and mainly self-organized criticality as SOC is a feature found in many complex systems. Publications in the Self-Organization Criticality category can then be sub-categorized.
- Contact Processes and Epidemiology on Adaptive Networks
- Game Theory on Adaptive Networks. This category includes social games at different levels of abstraction, evolutionary games

In the book *Dynamics On and Of Complex Networks* (Ganguly et al., 2009) the authors propose a simpler classification of scientific publications according to their traditional field of application:

- Biological Sciences
- Social Sciences
- Information Sciences

The authors, as stated in the subtitle of the book, restrict the classification to Biology, Computer Science and Social Sciences. Therefore, this classification is a subset of selected publications and their classification is according to the fields of the restriction.

3.7 Self-organized criticality

As seen, self-organisation is one aspect of particular interest in the study of adaptive networks. These networks are the result of feedback mechanism between topology and node/edge processes. Caldarelli and Garlaschelli (2009) propose that the combination of self-organised criticality and the theory of fractals are the ingredients of adaptive network models.

Self-organised criticality and adaptation in discrete dynamical networks deals with random boolean networks and random threshold networks that exhibit self-organised

criticality, showing that simple models can exhibit phase transitions from order to disorder and that global organisation is achieved by simple local rewiring rules (Rohlf and Bornholdt, 2009).

Their work on these types of random networks centres the dynamics of the networks into two co-evolutionary paradigms:

- An activity dependent rewiring of the relations between nodes where active nodes lose links and where frozen nodes acquire links.
- A correlation dependent rewiring of the relations between different nodes where nodes with correlated activity are connected and where nodes with uncorrelated activity are disconnected.

This co-evolutionary approach was shown to be very robust against noise (Rohlf and Bornholdt, 2009). The authors also state that this work still has some open questions that were not addressed. They notice the fact that the dynamics of these networks is not understood when near criticality for large numbers of nodes. They also acknowledge the need to couple these networks with external signals, mainly to provide network-environment interaction and then to study the networks evolution and finally they recognise that there is still the need for a discussion about application for these toy models.

Chapter 4

Q-analysis

Q-analysis[6] is a mathematical framework to formalize the structure of a relation between sets. It was developed and introduced to the social sciences by English mathematician Ronald Atkin and colleagues in the early 1970s and has been used as a research methodology in a diverse range of areas and contexts. Most of *Q*-analysis ideas are found in the initial reports and papers Atkin wrote during the development of the project "The Urban Structure Research Project" 1971-1974, in which Atkin wanted to answer a set of questions (Atkin et al., 1968, 1971; Atkin, 1972, 1974):

- How does the structure come into being?
- What are the components of the structure?
- How does a structure change with time?

Atkin et al. (1968) compare the pure mathematical perspective with the applied mathematical perspective to the question of defining the structure associated with a certain phenomenon. The authors assert that the two perspectives are interconnected and that significant advances in one are associated with advances in the other. For the authors it is a mistake for pure mathematics to be isolated from the physical observation of phenomena and it is equally wrong to think of applied mathematics without the study of the structures in abstract. In this work the authors propose that social sciences, namely sociology, must search for a mathematical language whose structure is easily mapped to the observations and concepts accepted by sociology. For this the authors reject languages based on real numbers, which the authors say present too much structure, and leads to a statistical treatment of the problems. They propose instead a language based in abstract algebraic

[6]The term *Q*-analysis first appears in Atkin's theoretical paper "From cohomology in physics to q-connectivity in social science" (Atkin, 1972, p.156)

topology as their first choice for these disciplines. They exemplify their choice with the application of Q-analysis, although it was not named Q-analysis in that paper.

Atkin (1972) writes in his review of his previous work on the search for a formulation for the physical sciences, that without violating accepted theories, it could be extended to the realm of social sciences. The author shows how the language of algebraic topology can be applied to social sciences as well to natural sciences. He shows how the q-connectivity and *simplicial complex* notions are developed as vehicles for the comprehension of co-cycles in physics.

In Atkin (1974) an application of Q-analysis to several fields, through a sequence of case studies, shows how some traditional phenomena that where treated under a classical perspective, could be studied and treated using an algebraic topology analysis. Namely, his work starts by introducing some ideas from group theory, following the discussion on its applicability to social problems. That work analyses the game of chess through Q-analysis, discusses the application of Q-analysis to art, and to the concepts of physical space connecting it with the notions from classic physics and quantic physics. The language of structures is also applied to the analysis of urban spaces and to decision making in university management. All these cases illustrate the descriptive power of Q-analysis as a language to understand the structure and parts in those complex systems.

In Freeman (1980) a distinction between Atkins 'backcloth' and 'traffic' on that backcloth is given:

> The algebraic topology of Atkin is based on a fundamental distinction between what he calls "backcloth" and the "traffic" on that backcloth. The backcloth is the structure of the "space" in which objects are located and events take place. The traffic consists of the objects and the events that are defined in terms of that backcloth.

Also, Gould and Johnson (1980) present the notion of "backcloth" and "traffic" when studying the content and structure of television flows saying that:

> Backcloth is a technical term, referring to a multidimensional structure formed from combinations of words chosen to describe the actual subject matter of television programs. Such structure is formed by a relation between a well-defined set of programs, say P, and a well-defined set of words, say B. (...)
>
> Traffic is another technical term, referring to the ways in which the actual

> subject matter of a particular television program may be treated, for example in a dramatic way, in a documentary fashion, for educational purposes, and so on. It may also refer to certain types of viewers, say children, watching a particular program, the amount of advertising revenue generated by a program, the number of people watching at a particular time, and so on. In brief, traffic as treatments or numbers has to exist on a backcloth: there must be some backcloth, some geometry, some structure there before the traffic can exist (Gould and Johnson, 1980, pp. 46-47).

Gould and Johnson (1980) oppose vehemently to the idea of 'classification' as a partition on a set and state that

> : 'partitional thinking' is totally inadequate to the task. Partitioning a set of television programs means taking a rich, highly-connected, multidimensional structure, and then severing most of the the relation in order to stuff these rich artifacts of contemporary culture into a series of little boxes. To be polite, such thinking is pathetic, (...) Gould and Johnson (1980, pp. 49-50).

The authors reject in this paper a traditionalist 'partitioning view' of science and claim that the 'Language of Structure' *is much more useful than pushing them through deterministic partitional machines called cluster algorithms, procedures which destroy most of the information contained in the sets* (Gould and Johnson, 1980, p. 50).

Johnson (1981b) wrote a reference paper for the standard definitions and notations on Q-analysis. His notations and definitions are adhered to in this thesis.

Johnson (1983), made a survey on Q-analysis where he discusses several aspects of this methodology at that date were not clearly established in scientific practice. The author starts by a clarification of the word *application* (in the sense of Q-analysis application) stating that it requires the investigation of backcloth-traffic relationships. For him instead of the use of *application* one would be better off with using the word *illustration*, when presenting the Q-analysis methodology through examples. The use of *application* should be reserved for industry application and not for scientific definitions.

Johnson also establishes that the theory of Q-analysis is a mixture of scientific imperatives, hypotheses and procedures whose application he characterises as follows[7]:

> "An application of the methodology of Q-analysis will involve <u>all</u> of the following:

[7] underlines by Johnson (1983)

1. set definition resulting in a (hierarchical) scheme of well defined sets
2. the collection of data as the observation of well defined relations between well defined sets
3. a connectivity analysis of the Dowker complexes of these relations viewed as (hierarchical) backcloth
4. the definition of traffic on the hierarchical backcloth
5. an explicit investigation of the relationship between the connectivity structure of the backcloth and the way it constrains the behaviour of the traffic it supports. In other words, an explicit consideration of the system kinematics and dynamics as backcloth-traffic interaction."

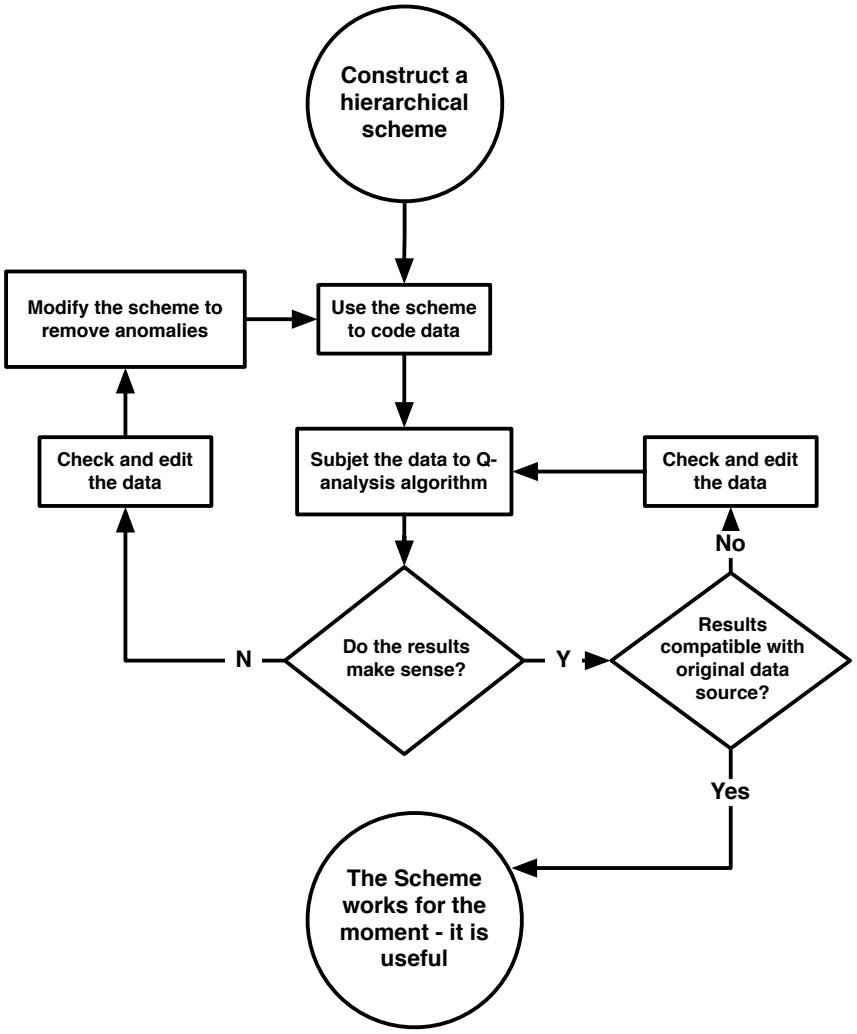


Figure 4.1: The iterative application of the *Q*-analysis algorithm (Johnson, 1983, p.8)

Johnson defends these conditions saying they are a set of interdependent scientific principles that represent an absolute minimum for research to be considered using *Q*-analysis, and in particular states that this definition supports the requirement that *if there's no traffic it is not* Q-*analysis* (Johnson, 1983, p. 5)

Johnson also states that *Q*-analysis is best applied through an iterative systematic process where one would be able to investigate a system by the careful examination of its parts as illustrated by fig 4.1; *'define your sets, collect relational data, subject data to the* Q-*analysis algorithm, do the results make sense (ideally in terms of the backcloth)?, if NO then go back to defining sets and collecting modified relational data.'* (Johnson, 1983, p. 6). The author recognises that up to 1983, the approach was heuristic and could only be judged on its results.

Johnson also states that *Q*-analysis is an attempt to move towards a scientific social science, and that this methodology contains some *'deep'* ideas for social science namely the notions of hard data and non-partitional hierarchy. Also the notion of backcloth-traffic is unique to *Q*-analysis and according to the author *'represents one of the major scientific innovations of this century'* (Johnson, 1983, p. 7)

On the applicability of *Q*-analysis to real policy making systems, Johnson claims that it is politically and ideological neutral as *Q*-analysis is a mathematical language of structures (Johnson, 1983, p. 13)

4.1 Advantages and limitations of *Q*-analysis

The use of *Q*-analysis as a descriptive language for structure has many advantages and some limitations.

Jacobson (2003) describes *Q*-analysis as "a geometrically oriented approach to exploring and representing structure in data, mapping relations among finite sets; draws on the topology of simplices and simplicial complexes; produces measures as well as graphs; and shares characteristics of both cluster analysis and network analysis".

The author also summarises some of *Q*-analysis advantages and problems as:

- Being descriptive rich
- Can include context and seeking behavior
- Laborious

- Quantifying is somewhat difficult

Beaumont and Gatrell (1982) present an overview of Q-analysis where they put in evidence the methodological and epistemological features that make Q-analysis attractive for mathematical description of structures and complex systems. The views are also shared by Atkin and others. See Beaumont and Gatrell (1982, pp. 49-51) for more details.

They summarize the advantages of Q-analysis in the following aspects:

4.1.1 An holistic approach

The concept of 'system' carries with it the notion that one is faced with interrelationships and interdependencies among the parts of the system, and that these are of crucial importance for the description and understanding of the system. It is not possible to understand the system behaviour solely by the inspection and analysis of its individual components and an additive analysis does not suffice in revealing the system in its entire dynamics and structure.

Q-analysis, by focusing on the structure of the simiplicial complex is holistic rather than atomistic, non-additive rather than additive and the structuralist perspective provides a language that can be easily adapted to a diverse range of problems.

4.1.2 A multidimensional approach

The multidmensional approach that Q-analysis employs does not present the caveat of other approaches where the relations are mapped to low dimensional spaces (for visualisation purposes, or for geodesic mappings in geography) where those high dimension level relations are distorted or even lost. Q-analysis retains this multidmensionality of the relations and the notion of 'space' is where one experiences through those relations in a contrast with a view of space as a simple 'container' of those relations.

4.1.3 A deterministic approach

Q-analysis is a mathematical language rather than a statistical technique and makes no reference to probability theory. This gives Q-analysis a powerful descriptive advantage in tackling problems rigorously for which statistical methods are inappropriate or non-existent. In any case Q-analysis does not exclude statistical treatment altogether and Atkin

(1974, 1981) discussed the circumstances under which the language of probability theory can be meaningfully applied.

4.1.4 A 'data-friedly' approach

Q-analysis relies on an unambiguous definition of the intervening sets and of the use of relations. This allows this language to be 'friendly' to the original data as no further transformations are needed. The authors contrast this with facto-analysis where data often needs transformation, calculation of correlation coefficients, extraction of factors, or factor rotation and standardisation of factor scores. At each of these steps the researcher has to make decisions and this might be followed by a numerical taxonomy. The authors state that this moves the analysis process further away from the original data. *Q*-analysis on the other hand, "might be said to let the data speak for themselves, and nothing is imposed on the data once the set and relations have been well defined" (Beaumont and Gatrell, 1982).

4.1.5 A 'scientific' approach

The lack of use of notions like theory, hypothesis or models in *Q*-analysis stresses the importance of data. The authors say that this might lead one to think of *Q*-analysis as anti-scientific. They, on the contrary, defend *Q*-analysis as scientific in a broader sense as *Q*-analysis provides a description of the data where as one usually thinks of science as providing 'explanations' of phenomena. The authors argue that this distinction between description and explanation is one that cannot be made simply and that there is a sense in which 'good description is explanation, for the intellectual content and meaning of the word explanation implies a description of relations between things' (Gould, 1980, p.171).

The authors state that this makes *Q*-analysis an appropriate methodology for 'soft' sciences in which 'hard' quantitative data may be more difficult to collect (Melville, 1976; Gould, 1980), going so far as to "draw attention to its place in critical social science as an emancipatory language" (Beaumont and Gatrell, 1982).

4.2 *Q*-analysis *'illustrations'* and similar work

Being defined as a tool to analyse structure of complex data, *Q*-analysis as been employed in many different contexts. Here some *illustrations*[8] are presented.

4.2.1 *Q*-analysis in social network analysis

Wasserman and Faust (1994) make a brief reference to *Q*-analysis in the field of social networks analysis and compare the sociomatrix, bipartite graph and hypergraph approaches as different representations of the affiliation matrix.

> "The two-mode sociomatrix, the bipartite graph, and the hypergraph are alternative representations of an affiliation matrix. All contain exactly the same information and thus any one can be derived from another. Each representation has some advantages. The sociomatrix is an effective way to present the information and is most useful for data analytic purposes. Representing the affiliation network as a bipartite graph highlights the connectivity of the network, and makes the indirect chains of connection more apparent. The subset representation in a hypergraph makes it possible to examine the network from the perspective of an individual actor or an individual event, since one actor's affiliation or an event's members are listed directly. However, the hypergraph and bipartite graph can be quite unwieldy when used to depict larger affiliation networks" (Wasserman and Faust, 1994).

4.2.2 Urban planning and architecture

Atkin et al. (1971) apply the language of algebraic topology (still without referring to it as *Q*-analysis) to the study of the commercial activities present in the city centre of Colchester, Essex, in the United Kingdom. They show how this method is adequate to express the existing structure, including both spatial aspects and non-spatial aspects of the city. In this study the authors make a comparative analysis of the *simplicial complex* of the actual city and of two restructuring plans for the city. With this exercise they show how the mathematical language of algebraic topology is adequate for the description and comparison of the different development plans. This is the first paper where the idea of

[8] Illustrations is the term preferred by Johnson (1983) when presenting the methodology through examples and applications

q-nearness is used in social systems research. The city is described by a group (L) of lozenges representing the streets of the city center, a group (A) representing the uses and activities found in that area, and a measure of distance between locations in the city through the street system. The authors defined four useful relations that illustrate the possible structures that one might find in this system:

- two lozenges, L_1 and L_2, are related if they overlap. In this relation λ_1, a crossing has four lozenges that overlap forming in this way a 3-simplex.
- a relation between the group of lozenges and the activities group given by λ_2. This defines two simplicial complexes $K(L, \lambda_2)$ and $K(A, \lambda_2^{-1})$. In the former, two lozenges are *q*-connected if they share some activity. In the latter activities are *q*-connected if they share the same lozenge.
- giving a length to each lozenge it is possible to define a relation λ_3 by saying that two lozenges L_i and L_k are connected if there exists a connection between them in $K(L, \lambda_1)$ whose distance is smaller than a parameter Δ. This defines the simplicial complex $K(L, \lambda_3, \Delta)$.
- giving a ranking (or partial ordination) to the ranking of group A, a λ_4 relation is defined in group A where A_i and A_k are related if the ranking difference between them is not bigger than a parameter ρ. This defines the simplicial complex $K(A, \lambda_4, \rho)$ where the ranking can be defined after economic consideration, empiracal data, or other important consideration.

This kind of structural analysis allowed, at the time of the study, to answer the needs that the city council had to understand the city's structure and the different renovation plans of the urban fabric that where being considered by the public authorities.

In the field of urban social areas Gatrell (1981) refers to *Q*-analysis as an appropriate framework in which to explore the urban social structures by providing a way to unravel the complexity of the structure of the urban social areas.

Also in the field of urban planning, Johnson (1981a) has shown how *q*-analysis can be applied to the description of road traffic systems. Traditional methods of describing the road transport system failed because they are based on a theory that is self-contradictory: the dynamic micro theory describing shock-wave behaviour in traffic queues is incompatible with the static macro theory based on equilibrium principles. The author shows how Q-analysis can be applied consistently to these two levels, the microlevel and the macrolevel. Also the paper makes use of the notion of *q*-transmission (Johnson, 1982), and illustrates

how the microlevel and the macrolevel can be combined in the more general context of land use. The authors shows that in this way it is possible to construct an integrated description of the land use-activity-transportation system (Johnson, 1981a).

4.2.3 Television programmes and media

In a work by Johnson (1978a) on television programmes, the author applies the Q-analysis methodology developed by R. H. Atkin in the field of television, by defining relations between programme types and time intervals of each of three British television channels, Anglia, BBC1, and BBC2. The simplicial complexes obtained from the construction of those relations are analysed by Q-analysis, and show how these simplicial complexes may form the natural backcloth for patterns of viewing and viewing demand.

In a research by Gould and Johnson (1980) research, the authors studied the flows of international television programs by application of Q-analysis. The study shows how Q-analysis can be suitable for understanding the cross-cultural impact of television, and help policy makers take adequate measures with the insights it provides.

4.2.4 Diagnosis of failure of large-scale systems

Q-analysis was used by Ishida et al. (1985) to build a failure diagnosis model of large-scale systems. A binary relation is defined by the set of units that are faulty and the set of measurements to detect a faulty unit. The authors study the relation between the simplicial complex topological properties and the diagnostic aspects. The authors verify that the capability of distinguishing a faulty unit by a given number of measurements is determined by examining the global connectivity structure of the simplicial complex. They observed that in real systems only restricted fault patterns are possible and that faulty units are functionally connected with each other, even in multiple fault situations.

4.2.5 Rule based systems

Duckstein et al. (1988) used Q-analysis as a tool to organise a pathological knowledge database, that is part of an expert system at the University of Arizona. Q-analysis was used in a module to obtain diagnostic assessment based on a rule-based system. Q-analysis is used to describe the relation summarised in a matrix representing four diagnostic categories and nineteen diagnostic clues of data concerning diagnosis of colonic cancer. The authors

show how *Q*-analysis is helpful to diagnosticians in defining and explaining the process they use in arriving at a diagnosis (Duckstein et al., 1988).

4.2.6 Multicriterion decision making

In Duckstein and Nobe (1997), the authors provide an overview of the existing and potential applications of *Q*-analysis in the design and analysis of expert systems in medical image processing: namely in the organisation of a histopathologic knowledge base. The authors also show how *Q*-analysis is applied to a multicriterion decision-making problem, using a method called multicriterion *Q*-analysis (MCQA). The authors also discuss the advantages and limitations of *Q*-analysis, with suggestions to future applications.

The authors explain that *Q*-analysis drawbacks include the following:

1. *Only qualitative measures of relations are obtained, in contrast with statistical analysis; however statistical analysis would require replications of homogeneous sets of data.*
2. *The mathematical theory behind the technique is not simple; however, a deep understanding of the theory is not required for application and correct interpretation of the results.*
3. *There are many indices that can be used, so conventions must be adopted for future comparisons of results.*
4. *The interpretation of results is not always straightforward* (Duckstein and Nobe, 1997, p. 423).

Also, the authors point out the main advantages of using *Q*-analysis:

1. *It is simple to use, requiring only "book-keeping" types of calculations.*
2. *It is flexible, there is no problem in changing slicing levels or criteria definitions.*
3. *It provides ordering on both direct and conjugate complexes; for example, q-levels, eccentricity, equivalence classes, and elements of the obstruction vector provide an order on the c_j's and d_i's.*
4. *It is applicable to several aspects of knowledge base analysis; for example, computerised histopathological image analysis.*
5. *It can be used in multicriterion decision-making methods (...), as well as many other problems involving the dynamic analysis of the structure of a binary relation* (Duckstein and Nobe, 1997, p. 424).

Chin et al. (1991) applied a multicriterion *Q*-analysis as a decision-making tool to choose automation projects for the printed-circuit-board manufacturing business. *Q*-analysis is used to group alternatives at different levels of satisfaction. A new eccentricity index can measure the relationships between alternatives. In addition, the conjugate complex, which transposes the alternative-criterion matrix, is analysed to rank the decision criteria and examine the relationships between them (Chin et al., 1991).

4.2.7 Communication content

Jacobson and Yan (1998) addressed the potential contribution of *Q*-analysis to the content analysis of print communication. The authors examined 27 news stories from the New York Times covering the "drug war" and compared the application of *Q*-analysis to other traditional methods and evaluated its merits. The authors conclude that *Q*-analysis is a promising approach to content research mainly because it allows for the mathematical expression of content properties beyond the reach of traditional content analysis methods (Jacobson and Yan, 1998).

The work by Jacobson and Yan (1998) also proposes that *Q*-analysis can be used in other contexts of text analysis besides finding structure in news, where much research indicates that news coverage exhibits patterns or structures. The authors also think that inferences about news values can be made possible by the use of *Q*-analysis and it allows also the possibility to relate news coverage to specific institutional practices within the media, and to specific audience contexts. Also *Q*-analysis could be applied to valence studies where the studies of backcloth and traffic could be employed to explore relations between topic categories and conflict coverage (Jacobson and Yan, 1998, p. 104).

Jacobson and Yan (1998, pp. 104-105) argue that the "most notable advantage of using *Q*-analysis is its flexibility in terms of the variety, and complexity of structures *Q*-analysis can represent. This flexibility may offer content analysts some useful tools for describing communication". They also state that *Q*-analysis avoids the use of data transformations and summary measures, stating that methods that retain data in their original form are sometimes advantageous.

The author also points out that *Q*-analysis presents two kinds of disadvantages, one procedural and the other substantive.

> "On the procedural side, the complexity that Q-analysis can sometimes reveal might at other times swamp analysis. This is partly true, perhaps,

> because standard computer software applications can only be used to perform a small part of the analysis. More importantly, the complexity of the structures revealed can themselves be formidable."
>
> "The major challenges of using Q-analysis, on the substantive side, are related to validating results. One area of difficulty involves estimating sampling or measurement error" (Jacobson and Yan, 1998, p. 105).

The authors also point out that in *Q*-analysis the definition of the slicing parameters is sometimes difficult, but point to the work by Johnson (1990) where rules for analysis were developed that are somewhat similar to those employed in expert systems.

Chapter 5

Community detection

Since the origin of graph theory with the work of Euler (1741), network science has evolved a considerable way. The XX century and early XXI century brought significant advances to the understanding of the properties of graphs. Such advances arise naturally from the fact that graphs are themselves an appropriate representation for the study of many complex systems, where the mapping of system properties to vertices and edges is many times straightforward.

In regular graphs, or lattices, all vertices present the same characteristics. One can picture a lattice as made of a repeating pattern, without any kind of randomness. In the opposite spectrum of graphs, one finds random graphs. In these graphs all connections are purely random and the probability of finding a connection between any two given vertices is given by a probability parameter p. These were the more interesting cases that were studied by Erdös and Rényi (1959, 1960, 1961)

Many complex systems can be mapped to network representations and usually the mapping does not correspond either to lattices or random graphs. These different graphs are representations that present several levels of structure and usually reveal high levels of heterogeneity. The degree distribution of the vertices is broad, many times following a power law distribution. This heterogeneity is also observed locally by observing that some groups of vertices concentrate many connections among themselves while having very few connections to other parts of the network. This is a property of real networks and shows that these networks are made of multiple smaller communities (Girvan and Newman, 2002). The algorithmic detection of these communities is also denominated module detection by some authors (Derenyi et al., 2005; Palla et al., 2005).

It is expected that the mapping of the complex system into a graph enables the depic-

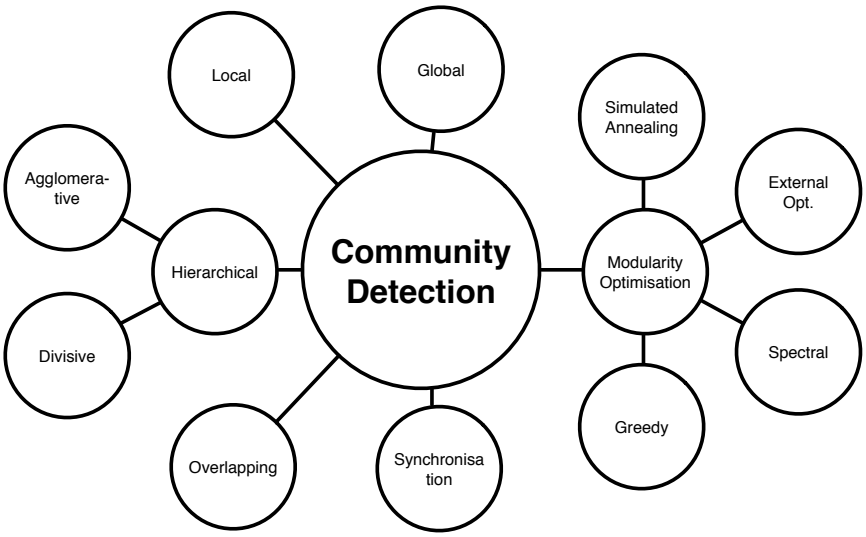


Figure 5.1: Community detection field

tion of those communities. An an example, the mapping of the personal relations of the inhabitants of a city or neighbourhood will naturally show the existence of communities representative of the different units present at different levels: family, work colleagues, friendship relations, etc... It has been precisely on the ways to determine these kinds of communities from the graph representation of the problem that network science has concentrated on in the recent years, trying to find commonalities, regularities, and eventually laws, that govern these real networks of complex systems.

Figure 5.1 shows the fields of interest in community detection in adaptive networks. It is important to highlight mainly the hierarchical agglomerative greedy modularity optimisation algorithms, that are used in the identification of communities resulting from the Q-analysis in Chapter 10.

Not being an objective of this work to exhaustively describe the advances that this new science produced recently in the field of community detection, it is still important here to recall the most important works. For a more advanced and comprehensive overview of the community detection subfield of network science, a recommend reading is Fortunato (2010) and the references therein. In that work a detailed review of community detection is made, together with their applicability limits. Using Fortunato's work as a guideline, this work follows the classification and nomenclature presented there.

Community detection importance is mainly due to the application potential of its results. The process can be thought of as a clustering of existing vertices into subgroups.

This is of particular interest for example for a web application where similar clients that are geographically near can help the results of a search engine, being served by nearby mirror sites. Another example is making a clustering of clients in online shopping according to their interest in similar products. This allows for the construction of better recommender systems. Besides these industry applications, knowing the role that each vertex plays inside each community can be very useful for example in the application of public policies, or organisational decisions inside a corporation. Central vertices in their community can have an active role in the control and spreading of information and community detection should reveal the areas of influence of these leaders. On the other hand, vertices that are on the periphery of communities but that connect to other communities in the network, will assume a role of interfaces to others, and will act as mediators, this could occur for example in business.

5.1 Measuring communities

The problem of finding communities in networks is one that is very transversal to many sciences, including social networks, computer networks, neuronal networks and metabolic and regulatory networks. These are found to naturally divide into communities or modules. Several approaches to finding the structure of these networks have been proposed and one that is very effective is via the optimisation of the quality function "modularity" (Newman and Girvan, 2004; Newman, 2006b). The notion of modularity as a measure of the structure of a network is that high values of modularity will reveal a network that has dense connections between nodes within a module but sparse connections between nodes in different modules.

Modularity index is defined as the fraction of edges that fall within the given modules minus the expected such fraction if edges were distributed at random and lies in the range $[-1/2, 1]$. Several ways to calculate modularity have been proposed but the most common version is one by the randomisation of the edges so as to preserve the degree of each vertex.

The formulation of modularity can be given in terms of the adjacency matrix and is useful for spectral optimisation algorithms (Newman, 2006b).

Suppose that a networks has n vertices and the entries A_{ij} of the matrix A represent the number of edges between vertices i and j (usually 0 or 1 if no multiple edges are allowed). If m is the total number of edges, given by $m = \frac{1}{2}\sum_i k_i$, where k_i is the degree

of vertex i. S_{ir} is defined to be 1 if vertex i belongs to group r and zero otherwise then

$$\delta(c_i, c_j) = \sum_r S_{ir} S_{jr} \tag{5.1}$$

and hence

$$Q = \frac{1}{4m} \sum_{ij} \sum_r \left[A_{ij} - \frac{k_i k_j}{2m} \right] S_{ir} S_{jr} = \frac{1}{4m} \mathrm{Tr}(\mathbf{S}^{\mathrm{T}} \mathbf{B} \mathbf{S}), \tag{5.2}$$

where $\mathbf{S}$ is the non-square matrix having elements S_{ir} and $\mathbf{B}$ is the modularity matrix which has elements

$$B_{ij} = A_{ij} - \frac{k_i k_j}{2m}. \tag{5.3}$$

If considering networks where the assignment of nodes is done just to two modules, it is possible to define alternatively $s_i = \pm 1$ to indicate the community to which node i belongs, leading to

$$Q = \frac{1}{4m} \sum_{ij} B_{ij} s_i s_j = \frac{1}{4m} \mathbf{s}^{\mathrm{T}} \mathbf{B} \mathbf{s}, \tag{5.4}$$

where $\mathbf{s}$ is the column vector with elements s_i.

5.1.1 Modularity resolution limit

As modularity compares the number of edges inside a module with the expected number of edges one would find a random network with the same number of nodes and with the same degree, this leads to problems in identifying small communities in large networks by modularity optimisation. This is due to the fact that the null model assumed (random edges with nodes with same degree) is unreasonable for large networks.

In the case of a large network the expected number of edges between two groups of nodes in this null model scenario can potentially be smaller then 1. In this situation a single edge between two modules would be interpreted as a strong sign of correlation between the two modules and would lead to an increase of modularity if joined by the optimisation algorithm, something that is naturally wrong. For this reason the optimisation of large networks would fail to resolve small communities. Fortunato and Barthelemy

(2006) showed that the probability that a module conceals well-defined substructure is the highest if the number of edges internal to the module is of the order of $\sqrt{2L}$ or smaller where L is the total number of links in the network.

5.2 Modularity optimisation based methods

In this work the modularity optimisation algorithm by Clauset et al. (2004) is used. This algorithm is a greedy agglomerative algorithm. This means that the algorithm starts with assignments of vertices to an equal number of modules each containing a single vertex. Then the algorithm proceeds iteratively by merging modules that maximise the increase of modularity at each step. This proceededs until only one community remains. The dendrogram resulting from the sequence of the merges can then be cut at a height corresponding to the highest value of modularity attained by the merging progression. This algorithm is less computationally expensive then other hierarchal algorithms and is therefore suitable to the study of community structure in large networks.

In any case it is important here to review some of the most recent advances in community detection algorithms.

5.2.1 Greedy methods

Among the algorithms based on modularity optimisation, it is important to describe some, mainly those based on greedy optimisation, simulated annealing, external optimisation, and spectral optimisation.

Modularity index Q (not to be confused with the Q in Q-analysis) was introduced by Newman and Girvan (2004) as the quality measure of the divisive clustering algorithm the authors presented at the time. This index quickly became very a important quality measure in several different community detection algorithms, mainly in those where the correct division of the network is achieved by the optimisation of the modularity index.

The combinatory nature of the problem of finding the best subdivision of the network makes it impossible to test all possible combinations for the higher modularity, even for networks of relatively small size. The number of calculations would be prohibitive and Brandes proves that the modularity optimisation problem is NP-Complete (Brandes et al., 2006).

In terms of greedy algorithms it is important to notice the hierarchical agglomerative one in which groups of vertices are successively merged forming communities in a way that increases the modularity of the resulting merge (Newman, 2004). This algorithm later leads to the algorithm of Clauset et al. (2004) that used the sparseness of the adjacency matrix of the graph to use a max-heaps structure to represent the graph. This lead to an algorithm based on modularity optimisation that could be used in large graphs (10^6 vertices being common).

This greedy algorithm usually presents the problem of generating large communities through the merging of smaller communities into larger ones. This is particularly evident when in the presence of communities with significantly different sizes. This led to improvements to the algorithm by different authors, namely Danon et al. (2006) that normalised the value of ΔQ of a particular merge on the fraction of incident edges in one of the communities; Wakita and Tsurumi (2007) proposed that instead of doing the merger according to the maximal value of ΔQ, the merge should be done on those two communities that maximised the multiplication of ΔQ by a factor called *consolidation factor*. This factor is maximal for communities of equal size. By using the *consolidation factor* the resulting algorithm presents a balance between modularity maximisation and community size and is a high performance algorithm (graphs with 10^7 vertices being commonly analysed). Another change was proposed by Schuetz and Caflisch (2008a,b) that suggested that at each iterations more than one pair of communities could be merged, obtaining in this several 'centres' around which communities would then form.

Some authors like Du et al. (2007), Pujol et al. (2006), Xiang et al. (2009), and Ye et al. (2008) sugested that better results can be obtained if the hierarchical agglomerative process is started from a reasonable structure instead of the isolated nodes. This reasonable structure has to be pre-calculated.

More recently, the Blondel et al. (2008) algorithm was introduced for the case where the graph edges carry weights. Initially all vertices are placed in different communities. The first step of the algorithm consists 1) in going through all vertices and for each calculating the weighted modularity gain that is obtained by placing that vertex in the community of one of its neighbours (alters), and choosing the community that maximises the increment of Q. After this first step, one obtains a first level partition. Next, 2) these communities are replaced by super-vertices and two super-vertices are connected if there exists at least one edge between two of its constituents. The two steps 1) and 2) are then iteratively

repeated until the value of Q cannot be improved by constructing higher hierarchical levels and higher level super-vertices. Although this algorithm performs better than the Clauset, Wakita, and Tsurumi algorithms, it has the problem of not being very precise in some situations. It is not clear what is the meaning of the intermediate levels, and the results are dependent on the order in which the vertices are processed in the first step of each iteration.

5.2.2 Simulated annealing

The use of simulated annealing in conjunction with modularity optimisation was first proposed by Guimerà et al. (2004) and it was first implemented by Guimerà and Nunes Amaral (2005). The technique combines two levels of actions on the graph: at a local level vertices are moved from one cluster to another randomly; at a global level the algorithm joins and splits communities of the graph. This method is very precise in obtaining the optimal value of modularity of the graph. It has the inconvenience of being a relatively slow process, and hence is really only applicable for graphs of up to 10^4 vertices.

5.2.3 External optimisation

External optimisation consists of a search heuristic first proposed by Boettcher and Percus (2001) that was applied with success to modularity optimisation by Duch and Arenas (2005). In this case the modularity was expressed as the sum of the contributions of the local modularity of all vertices of the network. The local modularity of each vertex is normalised by the degree of the vertex giving a fitness measure for this vertex. The algorithm is initialised with a random division of the vertices into two communities of equal size. At each iteration of the algorithm the vertex with the lowest fitness is moved to another community and the fitness is recalculated for all affected vertices. This process is repeated until it is not possible to increase the value of modularity Q. Each sub-graph is then submitted to the same process until no increase in the global value of Q is possible.

This algorithm presents the problem of being deterministic because it starts in the vertex with worse fitness and can become trapped in a local optimum. A solution to this is to include a probabilistic selection of the vertex to move based on their fitness distribution.

5.2.4 Spectral optimisation

The optimisation of modularity can be achieved by using the eigenvalues and eigenvectors of the modularity matrix B given by:

$$B_{ij} = A_{ij} - \frac{k_i k_j}{2m} \tag{5.5}$$

where A is the adjacency matrix, k_i and k_j are the degrees of vertices i and j, respectively, and m is the total number of edges in the network.

This modularity matrix leads to the modularity equation presented in 5.4. In that equation, the vector $\mathbf{s}$ can be decomposed on the basis of the eigenvectors $\mathbf{u}_i (i = 1, \ldots, n)$ of the modularity matrix B as $\mathbf{s} = \sum_i a_i \mathbf{u}_i$ with $a_i = \mathbf{u}_i^T \cdot \mathbf{s}$.

This leads to the definition of modularity as:

$$Q = \frac{1}{4m} \sum_i a_i \mathbf{u}_i^T \mathbf{B} \sum_j a_i \mathbf{u}_j = \frac{1}{4m} \sum_{i=1}^{n} (\mathbf{u}_i^T \cdot \mathbf{s})^2 \beta_i, \tag{5.6}$$

where β_i is the eigenvalue of $\mathbf{B}$ corresponding to the eigenvector $\mathbf{u}_i$. This form of the equation suggests that one can optimize modularity on bipartitions via spectral bisection, by replacing a Laplacian matrix with the modularity matrix $\mathbf{B}$ (Newman, 2006b,a).

This algorithm for finding communities has the advantage that it does not need the number of communities to be known in advance. The subsequent iterations of the division process of sub-communities stops when they do not lead to modularity increase (calculated from the adjacency matrix). The main drawback of the algorithm is that it gives best results for bisections, whereas it is less accurate when the number of communities is larger than two (Fortunato, 2010, p.31).

5.3 Detection of dynamic communities

Community detection has mainly been applied to static networks, or to timely snapshots of the networks in the case where they are dynamical, making it in reality a static analysis. In recent research some interest has been shown the study of dynamical communities, but the scarcity of annotated data with temporal sequences of events has made it difficult for these studies to progress (Fortunato, 2010). The main characteristics that one finds when analysing dynamical communities are moments in time that correspond to

important network events: community appearance, community expansion, community contraction, community division, community merging, and community disappearance. These phenomena are naturally associated with expressions commonly used in the life cycle of humans and in the literature many times one finds their anthropomorphic counterparts of birth, growth, death.

One of the first research made in this field was by Hopcroft et al. (2003) where they analysed temporal sequences of snapshots of the NEC CiteSeer citation networks between 1990 and 2011. The communities were identified by a hierarchical agglomerative algorithm where the similarity between nodes was calculated by the *cosine similarity* of the vectors describing the corresponding papers. This measure is very useful and is well used in information retrieval (Baeza-Yates and Ribeiro-Neto, 1999).

Recently Palla et al. (2007) made a systematic analysis of dynamical communities in two systems: 1) the network of telephone calls of a cell phone operator, for the period of two years; 2) a collaboration network between scientists, describing co-autorship of papers of *condensed matter physics* for the period of 142 months. The authors used the clique percolation method (Palla et al., 2005) and extended it for the analysis of the overlap of two consecutive snapshots of the evolving network. They concluded that the lifetime of a community is positively correlated with its size. They also showed that the evolution of the communities present different processes according to their size. While the composition of large communities is highly variable through time, relatively small communities are mainly static through their lifetime. Finally, they also showed that the probability that two communities merge increases with size and is higher would be expected than from the analysis of the size distribution of the communities. This provides evidence for the role of the faster dynamics observed in the bigger communities (Palla et al., 2007).

5.4 Detection of communities using synchronisation

Community detection algorithms have been inspired by many phenomena. One of such approaches considerers the phenomena of synchronisation (Pikovsky et al., 2001) as the underlying process of finding communities in graphs. If oscillators are placed at the networks' vertices, with initial random phases, then by subsequent nearest-neighbour interactions, oscillators that belong to the same community will synchronise faster, while oscillators in different communities will take longer to synchronise. This was first shown

by Arenas et al. (2006) where they used Kuramoto oscillators (Kuramoto, 1975, 1984). They placed a Kuramoto oscillator on each node of the network and by coupling each oscillator only to nodes in the neighbourhood of the oscillator. To reveal the effect of the synchronisation the authors introduced a local order parameter:

$$\rho_{ij}(t) = \langle \cos[\theta_i(t) - \theta_j(t)] \rangle \tag{5.7}$$

By inspection of this correlation matrix at a given time t, it is possible to distinguish groups of vertices that synchronise together. These groups can be identified by means of the *dynamic connectivity matrix* $D_t(T)$, which is a binary matrix obtained from $\rho(t)$ by thresholding its entries. By plotting the number of disconnected components found as a function of time, plateaus may appear at some time scales, revealing structural scales of the graph with robust communities (Fortunato, 2010, p.47).

Gfeller and De Los Rios (2008) investigated the possibility that coarse graining techniques could be applied as alternatives to large-scale simulations of complex systems. These systems are typically very large and difficult to simulate due to the great number of interactions present. The authors investigated how the dynamical properties of oscillator networks were affected when some nodes were merged together to form coarse-grained networks. They also showed that there are ways of grouping nodes that preserve crucial aspects of network dynamics, by considering some units as indistinguishable and merging them into a single node, thus giving rise to the concept of meta-populations. The authors investigated how "synchronisation in oscillator networks behaves when some nodes are merged together". For this purpose, they "introduced a Spectral Coarse Graining technique for the Laplacian matrix, which allowed them to know both how nodes should be merged, and which nodes should be merged in oscillator networks" (Gfeller and De Los Rios, 2008).

Chapter 6

Ant colony optimisation

As the problem of finding patterns in news is prone to combinatorial explosion, due to the vast number of elements and relations that might arise between elements, it is useful to refer to a family of algorithms that are capable of exploring vast search spaces in an efficient way. Research over hard NP-complete Combinatorial Optimisation Problems (COP's) has, in recent years, been focused on several robust bio-inspired meta-heuristics, like those involving Evolutionary Computation (EC) algorithmic paradigms. One particularly successful well-know meta-heuristic approach is based on Swarm Intelligence (SI), i.e., the self-organised stigmergic-based property of a complex system whereby the collective behaviours of (unsophisticated) entities interacting locally with their environment cause coherent functional global patterns to emerge. This line of research, known as Ant Colony Optimisation (ACO), uses a set of stochastic cooperating ant-like agents to find good solutions, using self-organised stigmergy as an indirect form of communication mediated by artificial pheromone. In this agents deposit pheromone-signs on the edges of the problem-related graph complex network. This encompasses a family of successful algorithmic variations such as: Ant Systems (AS), Ant Colony Systems (ACS), Max-Min Ant Systems (Max-Min AS) and Ant-Q (Gambardella and Dorigo, 1995; Dorigo and Gambardella, 1996; Bonabeau et al., 1999; Dorigo and Socha, 2006; Dorigo and Stützle, 2009).

Algorithm 1 ACO meta-heuristic

```
while end condition not satisfied do
    ants activity()
    pheromone evaporation()
    daemons actions()
end while
```

The ACO meta-heuristic was first introduced by Dorigo and Caro (1999); Dorigo et al. (1999) and is outlined in the Algorithm 1. The meta-heuristic was developed as a multi-agent approach to difficult combinatorial optimisation problems like the travelling salesman problem (TSP) and the quadratic assignment problem (QAP). The algorithms were inspired by the observation of real ant colonies. Ants are social insects that live in colonies and whose behaviour is directed at the survival of the colony as a whole rather than that of the single individual. An important aspect of the behaviour of ant colonies is their foraging behaviour, and how ants find the shortest paths between food sources and their nest. While walking from food sources to the nests, ants deposit a chemical substance called *pheromone*, forming a pheromone trail that evaporates through time. Other ants can sense (smell) this substance and this way find the shortest paths through the gradient concentrations of *pheromone* present in the environment (Dorigo et al., 1999).

This process is a kind of distributed optimisation mechanism to which each single ant gives only a small contribution. Although each ant is capable of finding a solution, it is the ensemble of ants, the colony working in parallel that is able to find the shortest path. This behaviour is an emergent property of the ant colony, mediated through the stigmergic variable *pheromone*.

This meta-heuristic is characterised by three steps, or kinds of activities. Each iteration starts with *ants activities* where ants making use of the local signals received from the environment to construct solutions for the problem. During this phase ants deposit pheromone on the paths that were chosen by them to construct the solution. Besides the ants' activity, ACO algorithms include a *pheromone evaporation* procedure that contributes as a negative feedback that allows the system to forget bad solutions over time therefore avoiding saturation. *Daemon actions* are activities that cannot be performed by single agents of the system, for example the additional deposition of pheromone on the globally best solution found, or the use of local search, 2-opt and 3-opt (Croes, 1958; Lin, 1965) in TSP, procedures to improve the quality of solutions found.

6.1 Traditional ant colony system

The traditional **ant colony system** (ACS) algorithm proposed by Dorigo and Gambardella (1996) is described in detail in this section as many anti-pheromone algorithms build upon it. In ACS a set of ants cooperate to find good solutions for the travelling

salesman problem by depositing pheromone in the edges of the TSP graph. This is a form of indirect communication of good solutions between ants. The formulation of the ACS algorithm is given by the equations 6.1 and 6.2.

State transition rule AS

$$p_k(r,s) = \begin{cases} \frac{\tau(r,s)\cdot[\eta(r,s)]^\beta}{\sum_{u\in J_k(r)}\tau(r,u)\cdot[\eta(r,u)]^\beta} & \text{if } s \in J_k(r) \\ 0 & \text{otherwise} \end{cases} \tag{6.1}$$

State Transition Rule in ACS

$$p_k(r,s) = \begin{cases} \arg\max_{u\in J(k)}\{\tau(r,u)\cdot[\eta(r,u)]^\beta\} & q < q_0 \\ eq.\ 6.1 & \text{otherwise} \end{cases} \tag{6.2}$$

The ACS algorithm is then an exploitation algorithm that tries to create solutions based on the best path available according to the ACS state transition rule 6.2 if a random number $q < q_0$, and acts as a random proportional exploration algorithm as in traditional Ant Systems given by the AS state transition rule 6.1 if $q \geq q_0$.

The ACS algorithm includes two pheromone updating periods: The first is denominated **local update** and occurs during the construction of solutions by each ant and acts as a way of diversifying the paths constructed by subsequent ants that traverse that node. The other, called **global updating**, is done at the end of each iteration (construction of one solution by each ant), and will deposit pheromone on only the best solution found.

The pheromone **global updating** rule is then defined as

$$\tau(r,s) \leftarrow (1-\alpha)\cdot\tau(r,s) + \alpha\Delta\tau(r,s) \tag{6.3}$$

where

$$\Delta\tau(r,s) = \begin{cases} (L_{gb})^{-1} & \text{if } (r,s) \in \text{global-best-tour} \\ 0 & \text{otherwise} \end{cases} \tag{6.4}$$

where α is the pheromone decay rate. The pheromone **local updating** is given by

$$\tau(r,s) \leftarrow (1-\rho)\cdot\tau(r,s) + \rho\Delta\tau(r,s) \tag{6.5}$$

where $0 < \rho < 1$ is a parameter for the local updating decay. The value of $\Delta\tau(r,s)$ can have different values but the authors determined that best results were obtained with $\Delta\tau(r,s) = \tau_0$ (Dorigo and Gambardella, 1996).

6.2 Anti-pheromone ant colony systems

Albeit being extremely successful, previous algorithms mostly rely on positive feedback, causing excessive algorithmic exploitation over the entire combinatorial search space. This is particularly evident over well known benchmarks such as the symmetrical Traveling Salesman Problem (TSP). In order to overcome this hard search space exploitation-exploration compromise, the present algorithmic approach follows a route coming from very recent biological findings which show the effectiveness of foraging networks were improved if different pheromones could also be used to repel foragers from unrewarding routes. Increasing empirical evidences for such a negative trail pheromone exists, deployed by Pharaoh's ants (*Monomorium pharaonis*) as a 'no entry' signal to mark unrewarding foraging paths (Montgomery and Randall, 2002; Robinson et al., 2005, 2007; Grüter et al., 2012).

Several algorithms make use of this anti-pheromone idea in an attempt to produce better travelling salesman problem (TSP) solutions:

The **subtractive anti-pheromone** (SAP) algorithm implements this anti-pheromone, not by using a true negative pheromone that is different from the positive pheromone, but by subtracting positive pheromone from the poor solutions found. This discourages subsequent generations of ants from experimenting with those paths in the construction of new solutions (Schoonderwoerd et al., 1996; Montgomery and Randall, 2002).

$$\tau(r,s) \leftarrow \tau(r,s)\cdot\rho' \qquad \forall(r,s) \in v_{worse} \tag{6.6}$$

This algorithm is similar to traditional ACS except that in the global pheromone update phase, the poorer solutions suffer a further evaporation given by equation 6.6, where ρ' is

the removal rate due to anti-pheromone and v_{worse} is the iteration worse tour.

The **preferential anti-pheromone** (PAP) algorithm was proposed following the work of Iredi et al. (2001) on an ant system for solving bi-criterion optimisation problems that uses two types of pheromone, one for each criterion. This idea is adapted in PAP using one pheromone for good solutions and another pheromone for poorer solutions through the inclusion of a parameter λ that defines the preference of a particular ant for the two different pheromones given by:

$$\lambda_{ant(k)} = \frac{k-1}{m-1} \tag{6.7}$$

where k is the index of the ant ($[1, \ldots, m]$) and m is the total number of ants. This implies that different ants will have different preferences for the two pheromones and while some will preferably explore paths in the vicinity of poorer solutions, others will preferentially search for paths in the vicinity of good solutions.

PAP reflects this balance of ants in equations 6.2 and 6.1 by replacing the pheromone term ($\tau(r,s)$) with a weighted term (on λ) over the two types of pheromones

PAP State transition rule AS

$$p_k(r,s) = \begin{cases} \frac{[\lambda\tau(r,s)+(1-\lambda)\tau\prime(r,s)]\cdot[\eta(r,s)]^\beta}{\sum_{u\in J_k(r)}[\lambda\tau(r,u)+(1-\lambda)\tau\prime(r,u)]\cdot[\eta(r,u)]^\beta} & \text{if } s \in J_k(r) \\ 0 & \text{otherwise} \end{cases} \tag{6.8}$$

State Transition Rule in ACS

$$p_k(r,s) = \begin{cases} \arg\max_{u\in J(k)}\{[\lambda\tau(r,u)+(1-\lambda)\tau\prime(r,u)]\cdot[\eta(r,u)]^\beta\} & q < q_0 \\ eq.\ 6.8 & \text{otherwise} \end{cases} \tag{6.9}$$

where $\tau\prime$ is the negative pheromone deposited.

A third anti-pheromone variant called **explorer ants** instead of using negative pheromones to mark poorer solutions, uses a small amount of ants as having preference for areas with little pheromone. These ants influence the environment by depositing pheromone in the same way as normal ants, only their preference for existing pheromone is reversed.

From these algorithms only PAP in reality uses two different pheromones but includes also different kinds of ants with their selectivity defined by the parameter λ. Both SAP and

explorer ants use only one kind of pheromone. In SAP the pheromone is decremented in poor solutions and in *explorer ants* a fraction of ants is deemed different from the majority of ants in the colony in their pheromone preferences.

In these approaches a two pheromone algorithm is introduced that does not change the classes of the ants present in the colony, keeping them equally selective to the pheromones.

6.3 Ant-based clustering algorithms

A revision of the main algorithms based in ant systems and stigmergy can be found in Dorigo et al. (2000). The stigmergic communication in these ant systems is usually implemented through a stigmergic variable. This variable can have different forms according to the application: it can be an artificial pheromone in the case of shortest path problems; it can be defined as energy levels in the case of foraging ants; it can be the level of consumer demand in the case of post delivery, or it can be the object distribution in a sub-dimensional space in the cases of data exploration and clustering (Dorigo et al., 2000).

Algorithms based on ant colonies are interesting because of some properties: the colony flexibility allows it to react to perturbations in the problem and eventually adapt to an entirely new problem, hence finding good solutions; the robustness in the execution of the tasks even if some of the agents fail or produce sub-optimal solutions; the decentralisation of the process avoids the need of any kind of central control; self-organisation as the solutions to the problems faced are emergent instead of pre-defined (Dorigo et al., 2000).

Here we will focus on some ant based algorithms aimed at the task of clustering.

Jafar and Sivakumar (2010) and Abul Hasan and Ramakrishnan (2011) wrote a global review of the main ant colony based algorithms for data clustering. The application of these techniques covers varied areas, but of particular interest for this work are those applications of ant systems in the area of recollection and treatment of textual documents. Insofar the following cases are particularly important:

Ramos and Merelo (2002) developed a novel strategy called ACLUSTER, for document clustering. The authors created an algorithm based on ants that move in a 2D space according to a probabilistic transition rule dependent on the deposited pheromone. In this way the authors avoid having ants visit random places of little interest. On the other hand if a cluster disappears, the pheromone will tend to evaporate in that place and a new solution will dynamically emerge. This perspective is interesting as the deposited pheromone acts

as the collective memory of the colony.

Handl and Meyer (2002) in *Improved Ant-Based Clustering and Sorting in a Document Retrieval Interface* applied a clustering technique based on ant behaviour as the core process of a web page retrieval algorithm. The main objective of their approach was to classify the retrieved documents according to their similarity. For this they employed ants with short range memory and different moving speeds. Also the ants could jump from one place to another distant one in certain conditions. Also, they applied an adaptive strategy to improve the results and the efficiency of the algorithm.

Abraham and Ramos (2003) applied the ACLUSTER algorithm in Web usage pattern finding and complemented the technique with genetic programming that allowed them to analyse visitor trends in the website.

Vizine et al. (2005) in *Text Document Classification Using Swarm Intelligence* use swam intelligence techniques for the automatic grouping of PDF files with potential application to Web document classification. The authors had to include two modifications to the standard ant-clustering algorithm that include the definition of a similarity degree between two texts and a cooling scheme controlled by a user parameter for the fine tunning of the convergence of the algorithm.

Chapter 7

Summary and research opportunities

In this part the main research in recent times on the main topics of interest for this theis: Text clusterings, Adaptive Networks, Q-analysis, Community Detection, and bio-inspired Ant Colony Optimisation algorithms were reviewed.

From the review the relevance of these topics for the understanding of complex adaptive systems, especially systems that are characterised by a high number of relations between their elements, where one wants to find patterns or similarity between them and that are continuously evolving has been shown. Online news publication presents these properties and is also a system which it is impossible to isolate in a lab for replication of the phenomena. This means that these languages and frameworks need to be particular suited to dealing with observation, description and analysis of systems that many times will not go under the same circumstances. Finding patterns in this volatile and high dimensional scenario requires hybrid approaches that exploit the structural properties of the data. There has been an attempt to combine some of these aspects over time, namely:

- Community detection and adaptive networks are just now starting to become important. The majority of community detection work has been done on finding communities on static networks, and the field is now turning its attention to Adaptive Networks and multilevel networks. It is believed that there is still a lot of research to be done in this particular conjunction of topics.
- Q-analysis was developed in the 1970s, 1980s and the interest of it faded away in more recent years. In any case it is clear that as new challenges emerge in the field of adaptive networks, it is necessary to find languages that describe these structures in ways that traditional networks could not. It is believed that it is possible to apply the Q-analysis framework to adaptive networks with the techniques of community

detection in novels ways. This will show new insights on the structure of online news that are not composed of binary relations but composed of high dimension n-ary relations between the vertices.

- Bio-inspired techniques present great advantages as heuristics that can solve hard combinatorial problems quickly without the need for the full exploration of the solution space. They are usually robust to change and can cope with dynamical problems. It is understood that developing new algorithms based on observations from biology allows for better solutions for the problem in hand as it is composed of many elements in a structure of interconnected relations. For this reason it is believed that heuristics that exploit the natural structural properties of the data can be helpful.

From what has been shown here there is the possibility to interconnect many of these fields of work into new algorithms, producing techniques that will give new insights into how complex networks of news behave. This thesis aims to combine these paradigms of Complex Systems studies. It aims at using the rich descriptive language of Q-analysis as a base for subsequent analysis of Adaptive Networks.

Part II shows how hybrid algorithms combining the connectivity of the published news and community finding (clustering) or bio-inspired algorithms can produce solutions that better explain this complex network of news by revealing the nature of their structural properties.

Part II

Hybrid connectivity based approaches

The second part of the thesis presents the algorithms and techniques proposed for the analysis of the structural properties of news published online under four different but inter-related approaches[9]. It starts with a short introduction to the dataset gathering and construction, after which the proposed hybrid methods are presented in subsequent chapters.

The datasets used in this research where collected over a period of 3 years from November 7th, 2009 through January 31, 2013. They consist of *html* files of the pages published online by the newspapers on their respective websites. This poses a big challenge due to a number of technical issues: different websites have different page structures; may add a publicity page before showing the news page or may setup a paywall and restrict access to certain news. For the purpose of collecting the dataset it was necessary to develop a crawling software called Theseus (see appendix A) that could deal with these various problems and that could do it in a reliable and permanent way. The pages were retrieved by fetching the published RSS[10] feeds. The RSS document usually includes full (although, sometimes only a summarised version is published in RSS) text of the news, plus metadata such as publishing dates, a permanent url, and authorship. From this RSS file that was fetched at predetermined intervals, the Theseus crawler fetches and stores the *html* webpage for later processing. Although a crawler, Theseus also included preprocessing and analysis modules, namely for the clustering techniques, presented in Chapters 9 and 10.

[9]These results were presented partly at the European Conference on Complex Systems 2009 in Warwick (Rodrigues and Louçã, 2009), at the European Conference on Complex Systems 2010 in Lisbon (Rodrigues, 2010), at the Rencontres interdisciplinaires sur les systemes complexes naturels et artificiels (Louçã, 2011), at the International Conference on Hybrid Artificial Intelligence Systems in Salamanca (Ramos et al., 2013), and at the European Conference on Complex Systems 2013 in Barcelona (Rodrigues, 2013)

[10]**RSS**: Rich Site Summary (originally RDF Site Summary, often dubbed Really Simple Syndication) is a family of web feed formats used to publish frequently updated works—such as blog entries, news headlines, audio, and video—in a standardised format.

Table II.1: Newspaper breakdown by country and language of journal

Newspaper	Website	Country	Language
El Pais	http://www.elpais.com/	Spain	Spanish
Le Monde	http://www.lemonde.fr/	France	French
Público	http://www.publico.pt/	Portugal	Portuguese
Bild	http://www.bild.de/	Germany	German
Folha de S. Paulo	http://www.folha.uol.com.br/	Brasil	Portuguese
The Guardian	http://www.guardian.co.uk/	United Kingdom	English
Los Angeles Times	http://www.latimes.com/	United States	English
Népszabadság	http://nol.hu/	Hungary	Hungarian
The New York Times	http://www.nytimes.com/	United States	English
Spiegel (international)	http://www.spiegel.de/international/	Germany	English
The Australian	http://www.theaustralian.com.au/	Australia	English

The versatility of RSS allowed the deployment of Theseus as the crawler of several projects of The Observatorium laboratory[11] including the gathering of generalist newspapers, social websites, science publications, and financial newspapers. This work focused on the generalist online newspapers publication. For that purpose representative journals were collected that were written in different languages and from different global geographical locations. Table II.1 shows a breakdown of the generalist papers crawled during the three years of acquisition of the dataset.

The use of newspapers written in different languages and in different geographical locations allowed for the detailed study of the features that these newspapers presented that are not language dependent. This allows for a higher level analysis of the complex structures that the news publishing activity exhibits.

Detaching the analysis from the particular semantic features of the corpora helps the the understanding of those properties that are intrinsic to the activity and not particular to any language or news topic.

From this collection of newspapers, two were chosen as illustrations for the approaches developed in this thesis. The two are the Portuguese newspaper **Público** and the British newspaper **The Guardian**.

The four approaches presented here are hybrid connectivity based in that they usually combine two techniques to classify the corpus of data as shown in table II.2. One of the techniques is always related to the underlying connectivity of the news documents, be it given by a construction of a dynamic adaptive network as in Chapter 8; given by the similarity matrix of feature vectors in the clustering of the timelines of *The Guardian* in

[11] Please visit http://www.theobservatorium.com/ for details on The Observatorium activities

Table II.2: Hybrid approaches developed during this thesis

Aspect	Chap. 8	Chap. 9	Chap. 10	Chap. 11
Variation of Information	✓			
Adaptive Network	✓			
Bags of Words		✓		
k-means Clustering		✓		
Modularity Clustering			✓	
Q-analysis			✓	✓
Bio Inspired				✓

Chapter 9; or by using the *Q*-analysis description of the connectivity of the news through the shared tags in Chapters 10 and 11. These connectivity based techniques are then coupled respectively with information based measures (8), clustering techniques (9 and 10), and bio inspired models (11), to give rise to novel hybrid approaches as sketched in table 1.1 of the Introduction.

The algorithms and techniques are tested on two particular datasets from the list of collected newspapers at *The Observatorium*. These are the **Público** and **The Guardian** newspapers that were collected daily for three years (2010, 2011 and 2012, see table II.1).

Chapter 8 presents the study of the **Público** newspaper. In this case it is shown how mutual information, more precisely a variant called Variation of Information (Meilă, 2007), can be used in conjunction with an adaptive network that is being built as news are produced to detect when clusters of topics disappear. This technique allows the identification of significant structures in the adaptive networks of news, where each document is inserted into a evolving network of other news and is connected to other news according to a measure of similarity (Rodrigues and Louçã, 2009; Rodrigues, 2010).

Chapter 9 explores the timeline of news published by **The Guardian** and clusters the news items based on a traditional *Bag of Words* model using *k*-means. An application was produced to show the timeline of the stories. The results were not satisfactory and this line of research was latter abandoned in favour of an alternative approach. Despite the disappointing result this chapter is still presented as *k*-means is usually considered a baseline clustering algorithm that is very straightforward to implement and to compare others to.

In a second approach in Chapter 10 the power of *Q*-analysis to extract the core of the news that are highly connected is shown. This is a novel procedure for automatic classification of a corpus of online news stories based on combining modularity optimisation techniques (Girvan and Newman, 2002; Newman, 2006b) and *Q*-analysis (Atkin

et al., 1968, 1971; Johnson, 1976). This chapter starts with the introduction of the overall system and then breaks it up in parts for deeper detailed comprehension of the structural properties of this news system. The main reasoning behind the idea presented in this section is that many news corpora are filled with *noise news*. *Noise news* are defined here as news items that are weakly related to other news for example because they are result of single non important events. It is claimed that this type of news will be sparsly connected to other news stories, mainly via some generic tags and can therefore be removed from the clustering phase that follows. The next step is then to cluster these cores and get highly correlated modules of news stories. This global process can be thought of as a filtering process that takes advantage of the structural properties existing in the data, without imposing a transformation or manipulating it in any particular way, and staying close to its underlying structure.

Taking advantage of the Q-analysis description, Chapter 11 then introduces a bio-inspired algorithm for finding Hamiltonian paths in the news structure of *The Guardian*. This approache uses the novel strategy of combining the eccentricity of the simplicial complex defined in chapter 10 with a novel double pheromone ant colony system that performs better than traditional ant colony systems in benchmark problems. The application of the algorithm to the news corpus shows that stories that share high dimensional faces are highly related and can be easily grouped together in a Hamiltonian path by the second order swarm intelligence (SOSI) algorithm.

Chapter 8

Topic monitoring with variation of information and dynamic networks

This chapter presents a system that aims to identify clusters of news from a stream of news. The idea is to derive a mechanism to handle incoming news in real time, associate it with existing news items and, in an automated way, group them into clusters of related news items. To solve this task a system was developed which combines an adaptive network of news that evolves in time as news items are added and old clusters of news removed, with a measure of the variation of information the system has between consecutive states. This conjugation of techniques allows for the identification of highly correlated clusters in news in an automated way[12].

This chapter also illustrates the use of the software Theseus in a concrete situation, that of automated classification of news stories from the Portuguese newspaper **Público**[13] using variation of information (Meilă, 2007; Rodrigues and Louçã, 2009; Rodrigues, 2010)

Large, real time text classification systems are becoming a popular topic. Here a method for automatically extracting correlated news from online media is presented. This method uses a dynamic similarity graph while using the variation of information as a measure to identify topics, lifespan and key terms. The method presented has the advantage of requiring no human intervention or training and having no pre-assigned categories because these emerge from the dynamics of the generated network.

[12]This chapter is based on the paper ***The Observatorium* – The structure of news: topic monitoring in online media with mutual information**, presented at the European Conference on Complex Systems 2010 (Rodrigues, 2010) following on the work previously presented at the European Conference on Complex Systems 2009 (Rodrigues and Louçã, 2009). A copy of these papers can be found at http://www.davidrodrigues.org/pdfs/rodriguesdavidms-eccs2010.pdf and http://www.davidrodrigues.org/pdfs/rodrigues_2009_eccs_mutual_information_sep_2009.pdf

[13]Público is the major printed and online Portuguese newspaper, available at http://www.publico.pt

8.1 Variation of information as a measure of topic change, as proposed by Meilă

The hybrid approach presented here uses variation of information (VI), which allows the measurement of the amount of information lost and gained when changing from partitioning C to partitioning C' of the same data set (Meilă, 2007). The following paragraphs present the VI measure, synthesising a detailed description by Meilă (2007). Considering one partitioning C, the probability that a node k belongs to cluster C_k is given by equation 8.1 where n_k is the number of nodes in cluster C_k and n is the total number of its elements.

$$P(k) = \frac{n_k}{n} \tag{8.1}$$

The uncertainty associated with the measure is the entropy of the variable $P(k)$

$$H(C) = -\sum_{k=1}^{K} P(k)\log(P(k)) \tag{8.2}$$

where $H(C)$ is the entropy associated with the partitioning C. This is always a non-negative value, and is zero when there's no uncertainty. The mutual information between two partitionings, C_k and C'_k, represented by $I(C_k, C'_k)$, means the information that one has about the other. The mutual information is given by the probability $P(k, k')$, representing the probability that a node belonging to the cluster C_k is in the cluster $C'_{k'}$.

$$P(k, k') = \frac{C_k \cap C'_k}{n} \tag{8.3}$$

Using expression 8.3, the mutual information $I(C, C')$ is defined as the mutual information associated with the two random variables k and k':

$$I(C, C') = \sum_{k=1}^{K} \sum_{k'=1}^{K'} P(k, k') \log \frac{P(k, k')}{P(k)P(k')} \tag{8.4}$$

Meilă (2007) proposed VI as a way to compare clusters:

$$VI(C, C') = H(C) + H(C') - 2I(C, C') \tag{8.5}$$

This measure is a metric because it is always non-negative, it is symmetric and it presents triangular inequality.

8.2 A new method for topic monitoring using Variation of Information and an Evolving Network

In this work the corpus consists of the news items gathered from the online Portuguese newspaper **Público**[14] from November 7, 2009 to January 25, 2010. The latest news RSS Feed were tracked and then the corresponding HTML files were downloaded.

After the collection stage, each HTML file has to be processed to remove duplicate files, HTML tags and artefact text. The problem with online media is that a webpage has textual information that is not pertinent to the topic extraction phase. It was also observed that a page's structure is not constant among all the retrieved pages, which prohibits the use of the HTML structure to retrieve the text as in Lin Lin and Ho (2002). Newspaper journals usually include also snippets of text from other stories that add to the difficulty of identifying the main corpus of text. Also because the HTML pages include several navigational links, these anchor text shows up as isolated words. Advertisements are also present and can pose problems.

To solve these issues, the text from each HTML file was extracted by employing the Text to Tag ratio proposed by Weninger and Hsu (2008) with a smooth factor of 1. This gives a good elimination of the text artefacts discussed above. Some word filtering was done to eliminate words shorter than three characters and longer than 20 (Cachopo, 2007). Some pre-processing techniques include other tasks, such as tokenization and stemming,

[14] http://www.publico.pt/

but these are language-dependent. In this case the objective of the system was to extract terms without previous knowledge of the text's language.

$$J(A, B) = \frac{A \cap B}{A \cup B} \tag{8.6}$$

Our method adds each text document to a graph, G, that is dynamically generated from the received text nodes. Assuming each node to have a certain life expectancy in this graph, the time to live (TTL), they were iteratively added to the graph and the Jaccard similarity (eq. 8.6) was computed between each node and all the previous nodes in the graph. Edges were then established from the added node to the previous nodes if their similarity was above a threshold j_{min}.

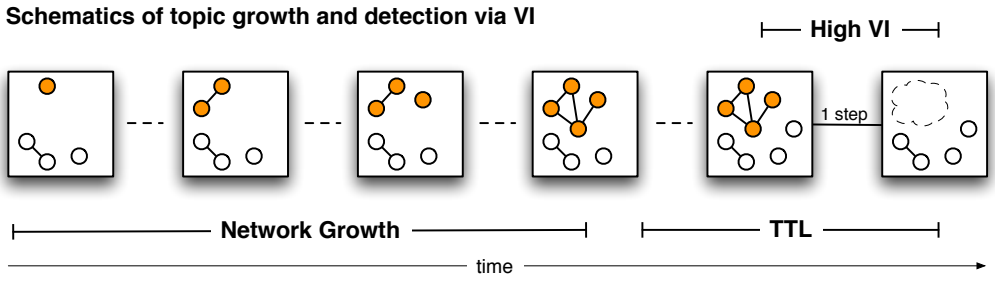


Figure 8.1: Schematics of network growth and variation of information on cluster deletion

Each time a node is added to graph G the TTL of the nodes that established connections to it is reset, but for other nodes this value is decremented. This purges old text nodes that do not receive new connections after some time determined by the TTL.

At each time step VI is computed from the previous graph to the new one $VI(G_{t-1}, G_t)$. From this value for which points VI exceeds a VI_{min} threshold can be determined. If the VI does exceed that threshold, it indicates the deletion of a reasonably large component of the graph. That also is the end of life for a topic in the timeline. From this information the topic's origin is traced by looking into that deleted component's oldest node, and the time span of each topic is also observed. It is also possible to process these components through some text extraction algorithm based on term frequency or summarisation technique. In this way it is possible to have highly focused summaries of the news clusters as soon as the topics fade away from the adaptive network. Although useful for *a posteriori* analysis this has the problem that variation of information only identifies them when they 'die'.

8.3 Results

The nodes of the graph were processed with $TTL = 100$, $j_{min} = 0.5$ and $VI_{min} = 0.5$. These parameters were determined empirically from a coarse search. These parameters are context dependent and can be adjusted online for each specific newspaper according to the results quality.

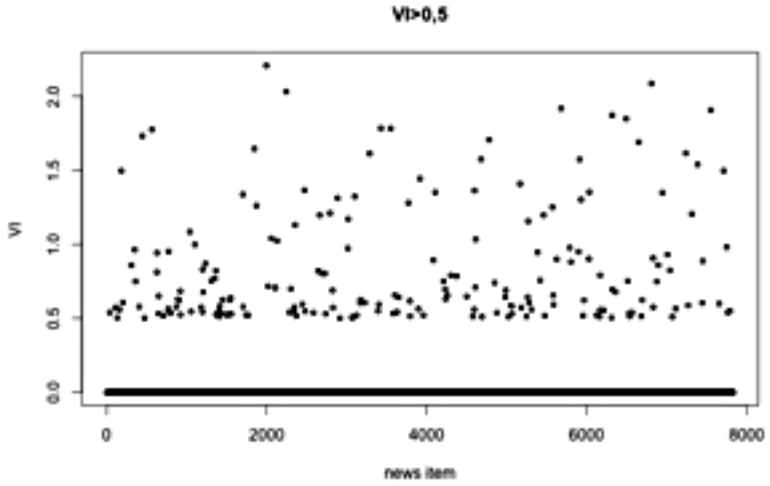


Figure 8.2: Evidence of topic deletion by tracking VI

Figure 8.2 presents the VI between the graphs of two consecutive time steps that exceed 0.5. Because VI is a metric, the points that present higher values represent greater changes in the graph's structure than do points that present lower values, which usually means the removal of an entire cluster of news from the graph. As an example, at time step 2092 $(VI = 2,2)$, the removal of a cluster of 37 texts (from late November 2009) is related to financial subjects. The main subject of those texts was the health financial scandal in the USA in November 2009.

By applying the proposed method 196 topics were identified during the analysis period.

Table 8.1: *Público* topic tracking results, av. lifespan and av. number news

n.topics	<Lifespan (/h)>	<Topic Size>
196	17.1	7.5

Those 196 topics had an average lifespan of 17.1 hours while the topic with the longest lifespan lived for 104.7 hours (approximately four days and nine hours). The results are in

accordance with what was expected from an online newspaper whose news concentrates mainly on daily topics with a few stories that "percolate" over several days.

8.4 Conclusions

This result shows that a method for simultaneously and automatically extracting topics classifying text is possible. This method has the advantage of requiring no prior knowledge or training. The use of VI allows for a fast method, based on information theory, for processing large volumes of data and allows the combination of network theory in the process of discovering online media's news structure.

The main advantage of this method is that it requires almost no human intervention. It has the limitation that the clusters are only identified when their time to live (TTL) goes to zero and those clusters disappear from the adaptive networks, thus forcing a big value of Variation of Information. This is clearly an obstacle for real time tracking, as sometimes news items persist and take a long time to disappear, although as we've seen the lifespan of a topic is approximately 17h. This is a clear indication that for the newspaper *Público* the news covered in this period is related to daily news.

This method is therefore practical for cases when one has past data with precise ordering of events and wants to analyse *a posteriori* the textual corpora.

Chapter 9

Clustering news: constructing timelines of news with *k*-means

While the hybrid approach shown in the previous chapter has some interesting qualities, it also has also some problems, namely that the adaptive network that evolves through time is an artificial construction in which the parameters TTL and threshold are arbitrarily chosen. This chapter aims to avoid this problem and to show how a traditional clustering approach behaves in explaining the structure of news.

9.1 A new algorithm for constructing timelines of news with *k-means*

For this a Bag-of-Words is used to build feature vectors representing documents that were then clustered with a *k*-means algorithm (MacQueen, 1967). *K*-means is usually considered as the baseline for clustering algorithms as it is straightforward and easy to implement despite being very computationally demanding for the task of text clustering due to the high dimensionality of the feature vectors. An application was developed for visualisation of the timelines of the news clusters formed by this strategy. The objective of this approach was to keep the time ordering of the events and therefore maintain a sense of narrative as the events where published by the newspaper.

A Bag-of-Words representation of documents is one where a fixed integer is assigned to each word in the document of the training set. This representation means that the total number of “features" is usually very large (larger than 100,000), and usually carries some properties problem that one needs to take into account when applying them:

- Bag-of-Words vectors throw away all information about word order. This can be partially mitigated by the use of n-grams[15].
- Most words do not appear in most documents; this means that the bag-of-words vectors are very sparse (most entries are zero).
- A small number of words appear in the majority of documents. These words tell almost nothing about the document topic. (*e.g.:* "the", "is", "of", etc.) This can be minimised if these stopwords are removed prior to vectorization.
- Most words are correlated with some but not all other words. Words tend to come in bursts

Usually it is useful to transform this occurrence vector representation of the documents into a term frequency (**tf**) based representation to normalise for different sized documents.

Another refinement on top of the **tf** representation is to downscale the weight of words that occur in many documents in the corpus and are therefore less informative than those that occur only in a smaller portion of the corpus. This process is called inverse document frequency (**idf**) and gives the well know downsizing **tf.idf** (Term Frequency times Inverse Document Frequency)

Algorithm 2 Pseudocode for the timeline of news algorithm

Require: HTML $corpus$ of News
  for all $document \in corpus$ **do**
    Convert $document$ to text
    Extract $sentences$ from $document$
    for all $sentence \in sentences$ **do**
      Clean $sentence$ of non-alphabetic symbols
      Extract $words$
      $BagOfWords \leftarrow stem(words)$
    end for
  end for
  for all $document \in corpus$ **do**
    Calculate feature vector of $document$ based on $tf.idf$
  end for
Require: $number\ of\ clusters$
  Compute clustering by $k - means(corpus, number\ of\ clusters)$.
  Launch visualisation application

For the analysis of the news publication of *The Guardian* through time a hierarchical

[15] **n-gram** is a symbol that replaces a contiguous sequence of **n** words in a document

clustering of the news was done. This hierarchical clustering was done on a corpus of feature vectors representing the documents in the corpus. The feature vector matrix was built using a Bag-of-Words and using the occurrence of each stemmed word weighted by term frequency, inverse document frequency (**tf.idf**).

For the application of the clustering with *k*-means it is necessary to define a similarity measure between documents. This was done by computing the cosine similarity between the feature vectors. Cosine similarity between two vectors u and v is given by

$$similarity(u, v) = \frac{u \cdot v}{||u||\ ||v||} \tag{9.1}$$

The algorithm no. 2 shows the pseudocode for the timeline analysis. The algorithm works by converting each of the HTML files into a text document, then proceeding to extract sentences from this document. From the sentences it cleans nonalphabetic symbols and builds a Bag-of-Words model by stemming them. Then after that all documents in the corpus are processed it computes feature vectors of each document based on **tf.idf**. This generates an incidence matrix that is clustered by the *k*-means algorithm. This clustering is then shown in the interactive visualisation application.

9.2 Interactive application for the analysis of timelines

The analysis results in an interactive application (part of Theseus, appendix A) where documents are partitioned into clusters through the *k*-means algorithm and are displayed chronologically for better visualisation of the relations. Also arcs are draw between those documents that have similarity above a user-defined threshold. This shows mainly the existence of relations intra-cluster, but also inter-cluster relations which shows that many stories could act as bridges between different clusters. Figures 9.1 and 9.2 illustrate a global view of the application and a detail, respectively. Each numbered label correspond to a news items published online by **The Guardian** during the period of 1 month.

In figures 9.1 and 9.2 it is clear that although most of the arcs are intra-cluster, as expected from the use an algorithm that tries to cluster by similarity, there is also some inter-cluster dependencies. This is a curious fact as it shows that the time relations between news can in fact be more localised than a hierarchical global method can explain.

These figures illustrate the phenomenon that some news stories, that are clustered

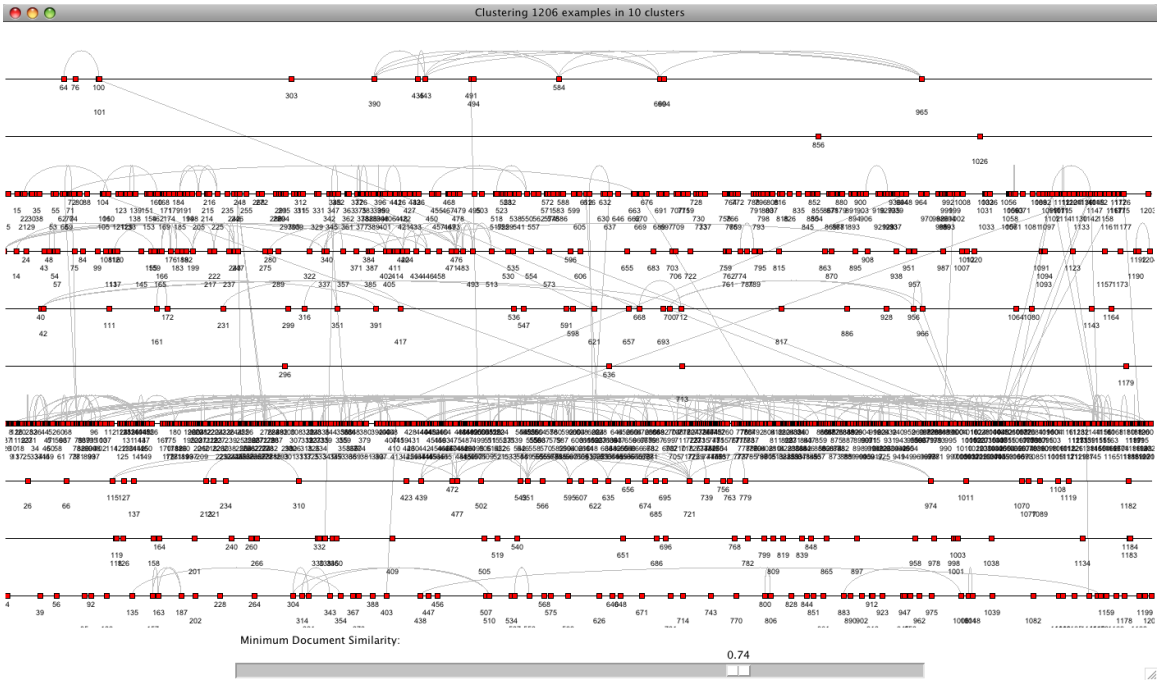


Figure 9.1: Global view of the application for the analysis of The Guardian timeline. Each line represents one of the clusters found by the *k*-means method.

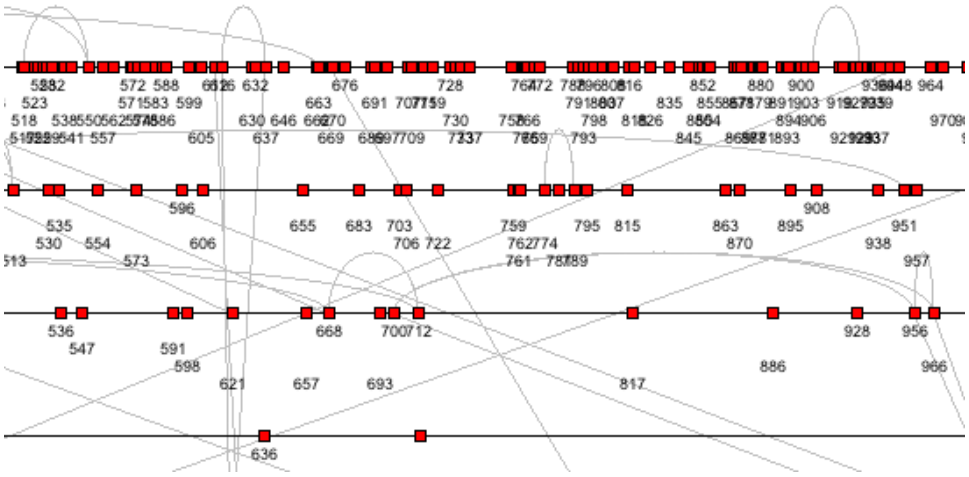


Figure 9.2: Detail of the time dependence arcs in the analysis of The Guardian timeline

together in the same cluster, are not sequentially connected as for example stories 700 and 712 in figure 9.2, and sometimes a story is highly related to future stories 'jumping' over other stories in that same cluster (see for example story 700 and 956). This shows that not all stories when processed by a global clustering algorithm have the same influence on future news stories and that the timeline gives practical information for relating the past of

news.

9.3 Remarks on constructing timelines with *k*-means

It is clear from the results of this exploratory work on construction of timelines, using a simple Bag-of-Words approach followed by *k*-means, clustering is not able to capture the fine structure and the relations of the news.

Some of the criticisms about the use of *k*-means clustering is that a euclidean distance between the feature vectors is used as a metric and in the context of textual analysis this might not represent a meaningful measure. Different words have different importances in different contexts. So a simple statistical representation of the documents as **tf.idf** is not able to translate the richness of the relations existing in written documents.

The *k*-means algorithm can be considered as a baseline for clustering techniques as it is straightforward and easy to implement, but suffers from high computational cost, which makes it less appealing when dealing with large datasets. In this case the algorithm was used for 1 month of news and could not be expanded any further because of the growing space of the feature vectors. As the results proved not to be of great interest mainly because of the need to define *k* as a parameter this line of research was abandoned in favour of others that could automatically find the correct number of clusters in a corpus of documents.

Also the *k*-means algorithm is prone to giving partitions of the dataset that are non-optimal, meaning that the algorithm can converge to local minima producing counter intuitive results.

Another problem with the use of *k*-means as a clustering technique is that the use of the number of clusters as an input parameter is problematic in the context of the corpus of data studied. Although the timeline application proved useful for aspects related to ordering of the news, it also showed a great dependence on the number of initial clusters, initial cluster centres and similarity threshold. This technique alone, although informative, has problems as it requires *a priori* knowledge of the number of clusters. This is usually an unknown in the news published by newspapers and the application of rules of thumb (e.g. the number of sections of a newspaper, or the square root of half the number of items) will not allow for a proper explanation of the resulting clustering. The next chapter will suggest an alternative method to overcome the problems identified here. It will try to compose a

novel strategy to cluster the news using the connectivity of the stories based on *Q*-analysis and the high dimensionality of the edges between news stories.

Chapter 10

Clustering news: finding communities with *Q*-analysis filtering

The main problem detected in the work of the previous chapters is that all news is included in the solution of the algorithm, be it in the form of nodes in the Adaptive Network, or as a node in the constructed timeline of the clusters. This means that the algorithm is forced to classify a story into a label, or category, even if it is just a singularity, or an isolated event.

It is important to clarify why these singularities are called *Noise News* in this application. **Noise news** is a news story that is published by an online newspaper but that does not seem to have any relevance or connectivity to other stories in the newspaper. These are still connected to other stories because they might be under some hierarchical umbrella like *local news*, or *world news*, but are not related in any other ways to the rest of the news. This makes them *noise news* in the sense of identifying clusters of news.

This is a broad definition, but helps in understanding the main advantage of the following technique, that is to find topics of news, that are free of singularities, or noise news, that do not bring any added value to the topic in question. And do it in a way that is automated and does not require human intervention or any semantic knowledge of the system studied.

This section aims at explaining how to overcome this problem by the application of *Q*-analysis, helping to filter noise news and improving the quality of the resulting clusters when one attempts to find cohesive groups of stories that would form a news topic[16].

[16]Part of this work as been presented at the European Conference on Complex Systems 2013 in Barcelona (Rodrigues, 2013). Pdf copy is available at http://www.davidrodrigues.org/pdfs/2013/david-rodrigues-eccs13.pdf

10.1 Clustering news with the Fast Community Algorithm by Clauset el of. (2004)

The Guardian[17] classifies every news item published with a set of metadata that can be used for clustering information. The two most interesting metadata fields are the **section** and the **tag** metadata. Each document has one **section** field corresponding to the section of the newspaper where the story was published and one or several **tag** fields that the journalist / editor chose to characterise that particular story. This proposal takes advantage of this human labelling to characterise the structure of the network created by all the news from **The Guardian**.

A first approach was to consider that two stories were connected if they shared at least one tag among them. From the resulting graph the communities present int the graph can be found by using modularity based methods. We will use the data from the month of November 2011 for this illustration.

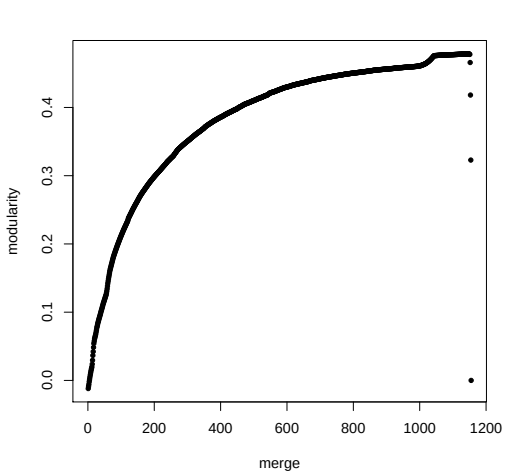
For the purpose of using a community finding algorithm a fast greedy algorithm proposed by Clauset et al. (2004) was chosen. This algorithm is a hierarchical agglomerative algorithm for detection of community structure that is very fast. The algorithms runs in $O(md \log n)$ for a network with n vertices, m edges and where d is the depth of the dendrogram. For networks that are hierarchical this means that $d \sim \log n$ and if the networks are also sparse, then $n \sim m$ making the running time essentially linear at $O(n \log^2 n)$. This allows for a quick overview of the community properties present in the graph of tags[18].

The fast greedy algorithm is an agglomerative algorithm and by plotting the modularity of the communities found at each merge one can find at which iteration of the merging process the maximum value of modularity is obtained. This is the point at which one selects the cut in the dendrogram and identifies the communities present.

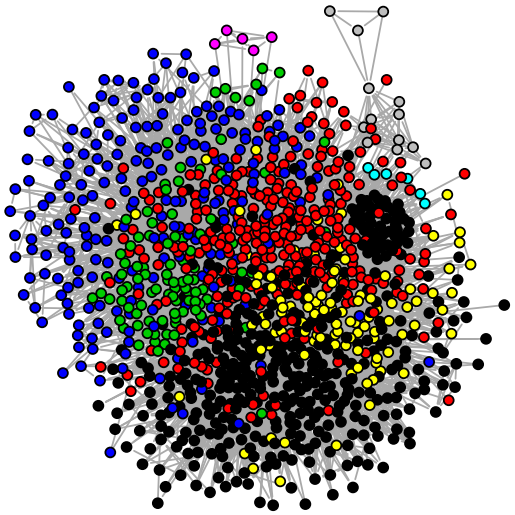
It can be observed in figure 10.1a that the maximum value of modularity (0.48) is observed for the merge number 1146 corresponding to 9 communities left. These would decrease the value of modularity if merged any further. The communities found can be visualised in figure 10.1b, and it is clear that they are not easily identifiable. There is to much noise, mainly because many labels used by the Guardian authors/editors are too general. Labels like 'World News' or 'Sport' abound and these labels can reveal very little

[17] http://www.guardian.co.uk/

[18] The processing is done in R (R Development Core Team, 2011) using the package *igraph* (Csardi and Nepusz, 2006) that has implemented many useful algorithms for graph manipulation.



(a) Modularity at different steps of the merge process



(b) News Tags network colored by communities

Figure 10.1: Modularity of the clustering fast greedy algorithm by (Clauset et al., 2004) and resulting communities.

of the underlying structure of the real topics.

Table 10.1: Cluster size distribution

id	1	2	3	4	5	6	7	8	9
items	363	303	96	221	6	5	102	13	46

This noisy classification can be seen in the detailed in the count of each of the communities presented in table 10.1.

This clustering reveals large components with more than 300 news items (cluster 1 and 2), meaning that probably a division into more components would be of greater interest and that these large clusters do not capture the fine structure of the news. This phenomenon was also observed in the previous chapter with the clustering through k-means. It is suspected that noise stories are connecting the system into a highly dense cluster and masking the underlying structure of the news. Because of this a new approach is needed, one that can give effective insights into the news structure.

Applying a modularity algorithm alone to the network of tags does not reveal enough structure to make the information useful in partitioning the news graph into useful subcomponents. The network has a density of edges of 2.04% and a single component with an average path length of 2.32 and a diameter of 18. But this does not provide insightful

information as to which topics where really being discussed during that period.

To understand this a new analysis was perfumed that explored how *Q*-analysis could be used to filter the noise news that abound in online media due to the broad labelling practice.

10.2 A new approach for clustering news using *Q*-analysis

The Guardian news from 2010, 2011 and 2012 where collected and a monthly analysis by *Q*-analysis was made. The resulting networks were studied in terms of edge density, number of clusters, modularity of the resulting clustering, maximum cluster size, fraction of vertices present in the resulting graph, and assortativity.

By plotting these monthly measures as a function of the connectivity (value of q) one can grasp the structural characteristics of this publication.

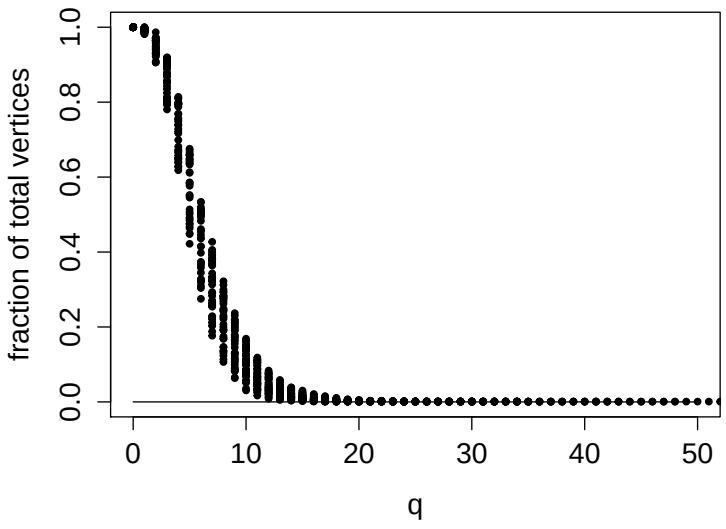


Figure 10.2: Fraction of vertices in the resulting graphs as a function of q

The first aspect to take into consideration when applying this method is the number of vertices (news) that is filtered out when considering the connectivity of the system. One can put this in other terms by considering the fraction of vertices that is kept in the structure by filtering out all that do not have at least connectivity q. This is represented in figure 10.2 where it is observed that 80% of news are connected for q=3. It is observed that very few

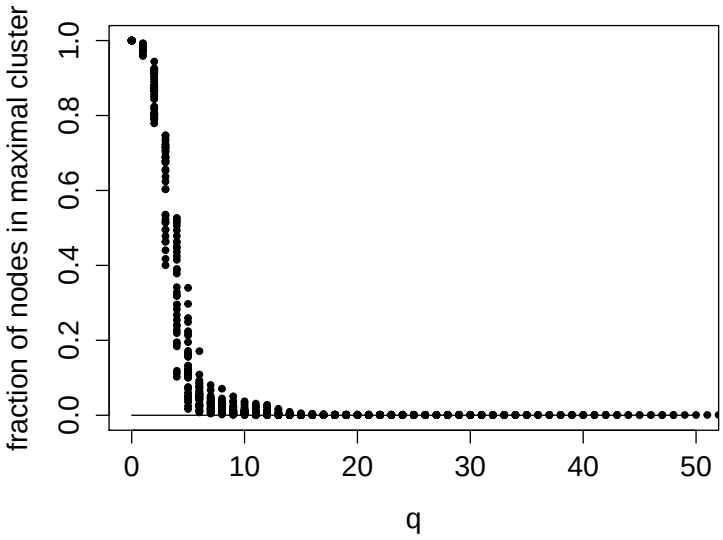


Figure 10.3: Fraction of nodes in the maximal cluster as a function of q

stories are retained if one increases the connectivity threshold of the resulting structure. For a connectivity of $q = 10$ the resulting structures retain less than 20% of the original nodes. This sharp drop of the fraction of retained nodes is indicative that the majority of the news items are connected through low dimensional faces ($q = 3, 4, 5$). This is also observed by looking at size of the biggest cluster resulting from the modularity detection algorithm.

In figure 10.3 it is observed that for low values of connectivity there is a giant component that captures most of the nodes in the structure, but for values of $q >= 5$ the fraction of nodes present in the maximal cluster drops substantially and the giant component disappears. This transition from giant component to the more regular components is also indicative of a drastic change in the connectivity of the resulting structures after filtering the structure of news of low connectivity news.

The low value at which this transition is observed provides evidence for the low connectivity of different stories and might be a reflection of the way the tagging mechanism works. Maybe journalists tend to give each story a few specific keywords instead of more general ones that would be broader in meaning (although increasing the connectivity of the news story).

While figure 10.3 represents the fraction of nodes relatively to the total number of

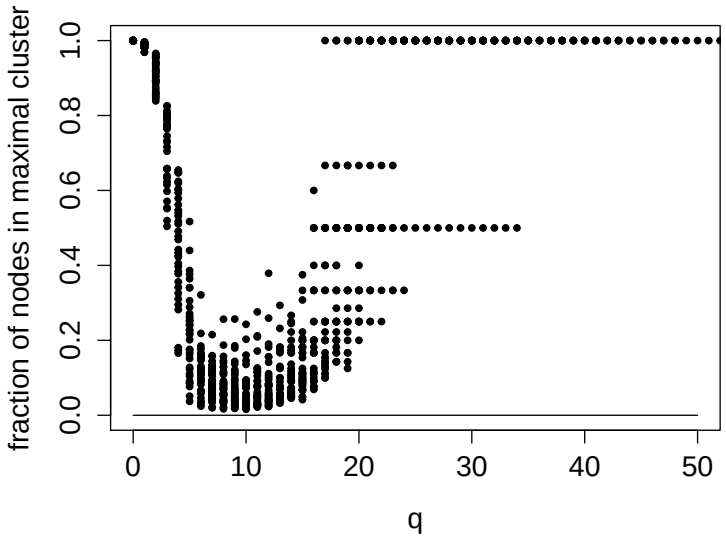


Figure 10.4: Fraction of nodes in the maximal cluster relatively to the number of nodes in that particular graph

nodes in analysis, a better understanding of the system can be drawn from the fraction of nodes present in the maximal component in relation to the total number of nodes of that q-graph.

This can be seen in figure 10.4 where for small values of q ($q < 5$) the giant component dominates. For higher values of q ($q > 14$) one start's to see that the giant component starts to dominate again. In these cases this is due mainly to the fact that the giant component is very small (figure 10.2) and unique (figure 10.5).

For the purpose of filtering content that is not highly connected and therefore unrelated to other news it is important to choose a value of q as a cutoff to optimise some properties of the resulting induced graph. These properties include the retention of a majority of nodes, high number of clusters, relatively small size of the maximal cluster and high modularity from the resulting graph. From figure 10.6 it is clear, after processing each induced graph with the agglomerative hierarchical algorithm by Clauset et al. (2004), that the resulting induced graph presents high values of modularity for $q > 3$. According to Newman, a network that presents modularity above 0.3 is considered to have some sort of community structure, meaning that the density of edges intra communities is much higher than that the density of the edges inter communities (obviously, when compared against a null model where a network is constructed randomly with the same number of edges/nodes

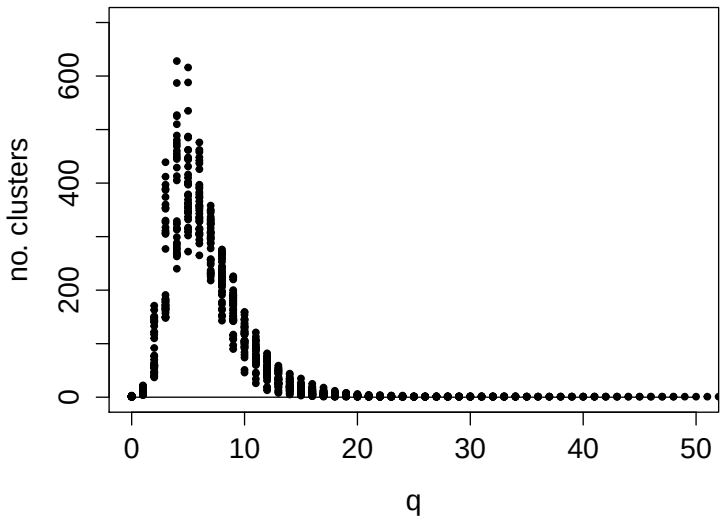


Figure 10.5: No. of clusters as a function of q

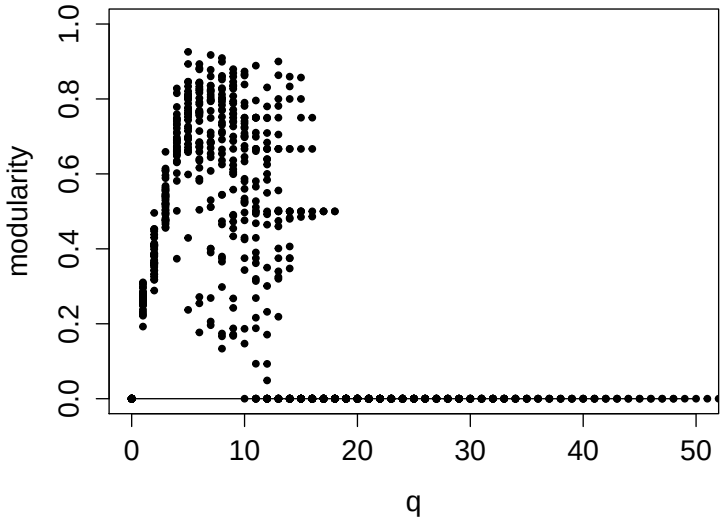


Figure 10.6: Modularity of the induced graph as a function of q

and degree distribution). In the following case we will look into more detail to the month of November 2011 in more detail. A detailed breakdown of the topic clusters is presented in appendix B.

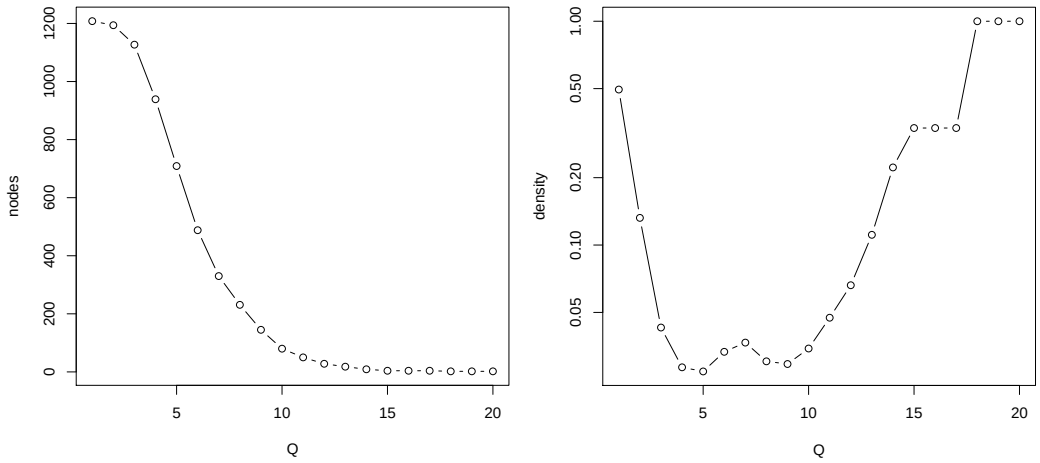
Algorithm 3 Pseudocode for the automated news clustering and filtering algorithm

Require: $RSS\ feed$
  /*Initialization*/
  for all $news\ item \in RSS\ feed$ **do**
    Fetch HTML page
    Convert to Text
    Extract Tags
    Generate Simplicial Complex (each news item is an actor, each tag is a feature)
    for all values of q **do**
      Extract graph G_i with simplexes that are at least q-near with other simplex $Q \geq q$
      Use modularity optimisation to compute modules in graph G_i
    end for
    Rank graphs (G_i) according to modularity index obtained from clustering algorithms
  end for

Each news item is tagged by The Guardian. This leads to the construction of an overview of the published stories. For this a description of the system was done with a *Q-analysis* of the Documents-Tags bipartite relation. In *Q*-analysis (Johnson, 1981b, 2005, 2006a) two nodes are connected by a link if they share at least $(q + 1)$ common attributes. In this case two news items are connected if they share at least $q + 1$ tags in common. The resulting networks constructed in this way are not tag networks as in the previous section, but document networks where two documents are connected if they share at least *q* tags in common. The November 2011 documents network was studied in detail for *q* in the range $[1..20]$. A pseudocode for the algorithm of automatically processing (filtering and clustering) the monthly news published by **The Guardian** is outlined in algorithm 3.

For each resulting *q*-graph several network properties were measured: number of nodes, number of edges, number of components of the graph, density of the graph, average clustering of the graph, degree assortativity, graph clique number and graph number of cliques.

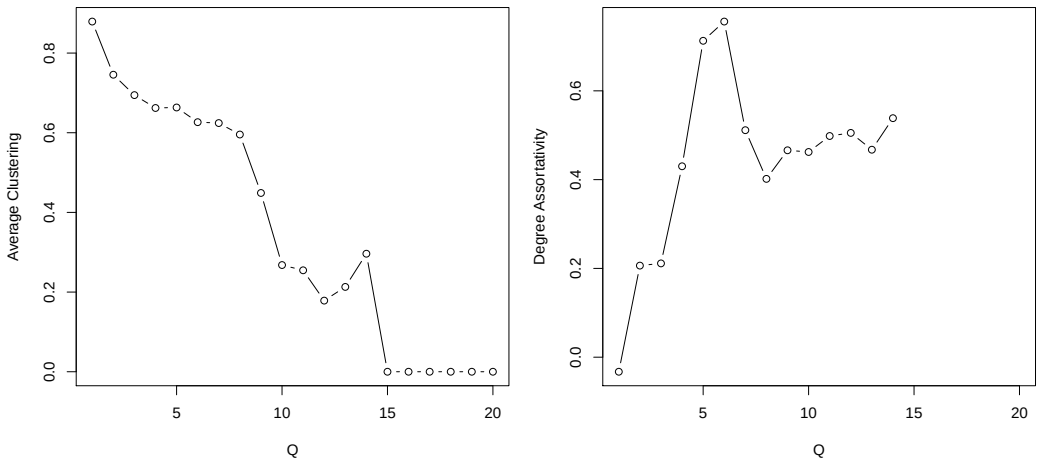
Not all these properties have the same importance for the discussion of community finding in networks, but for this case it is important to define what is understood by community. A community of a graph is a subset of nodes of that graph where link density is higher among the nodes of the subset than the inter-module connections. This definition



(a) Number of nodes **(b)** Edge density

Figure 10.7: Number of nodes present and edge density as function of q

by itself is the basis of modularity based approaches to finding communities in graphs. On the other hand it is clear from the previous section that using modularity alone does not give an interesting partitioning of the network. By applying Q-analysis to the bipartite graph, one can find subgraphs of the initial graph that are more 'modular' than the initial graph. This implies a tradeoff between completeness of the graph, as q increases, the number of total nodes q-connected decreases sharply as seen in figure 10.7a.

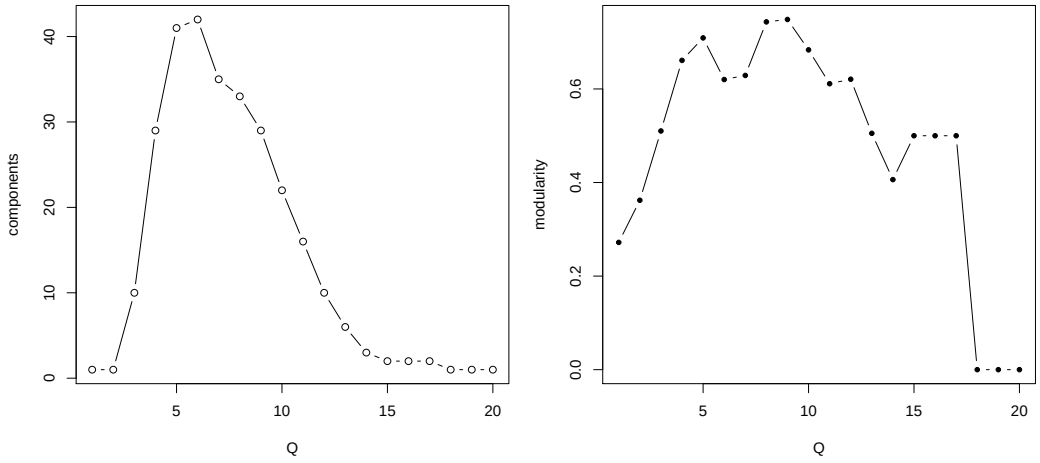


(a) Average clustering **(b)** Degree assortativity

Figure 10.8: Average Clustering and Degree assortativity of the subgraph as function of Q

For identifying networks with high structure that are highly modular, the properties that are of most importance are edge density (usually required to be low) figure 10.7b,

clustering (high) figure 10.8a, assortativity (high) figure 10.8b, and if by increasing the value of q one obtains disjoint graphs (graphs with more than one component), the number of components of a graph might also be a good indicative measure of the existence of structure, figure 10.9a.



(a) Number of components **(b)** Modularity

Figure 10.9: No. of components and Modularity index as a function of Q as calculated by the fast greedy algorithm of Clauset et al. (2004)

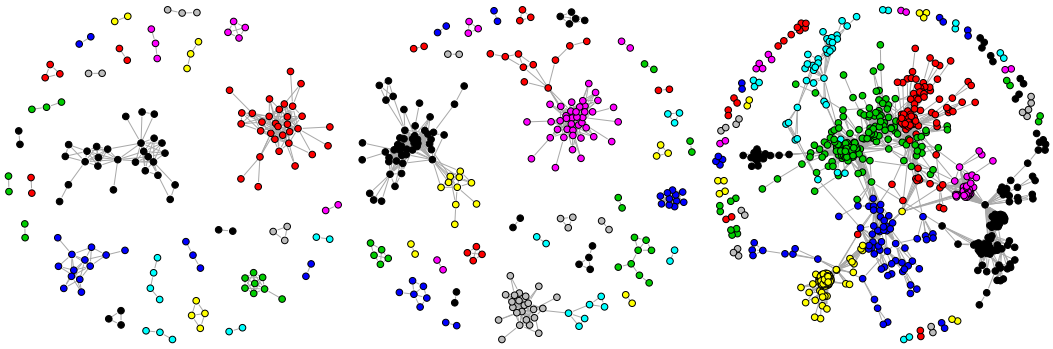
This process of dividing the graph might find parallelism with other hierarchical methods like the ones applied by (Girvan and Newman, 2002) where edges are removed by edge betweenness until components are separated and modularity index is then used to determine the point of the dendrogram to cut the graph.

In this case for each resulting q-connected graph, the value of modularity index is calculated by applying a greedy fast community finding algorithm (Clauset et al., 2004). Figure 10.9b show clearly that for low values of q the modularity of the resulting graph is small and it increase as one increases the cutoff value of q. This is in agreement with the global observation for the entire period of the study shown in figure 10.6.

10.3 Maximum Modularity of the *Q*-analysis graphs

Analysing the maximum modularity for each of these graphs it is observed that a maximum value of modularity is obtained for q=9, followed by q=8 and q=5 (figure 10.9b).

From the figures 10.10a, 10.10b, and 10.10c it is clear that increasing the value of **Q** gives better values of the modularity index but at the expense of discarding many nodes of



(a) Q=9, μ =0.75 **(b)** Q=8, μ =0.74 **(c)** Q=5, μ =0.71

Figure 10.10: Guardian news, colored by community:

the network (figure 10.7a). This means that $Q = 5$ seems to be a better choice as it still has a high value of modularity ($\mu = 0.71$) and keeps more nodes in the document. This is at the moment a task that requires human intervention, but one could for example devise a simple strategy to choose the cutoff value of q like 'increase q until 10% of nodes are removed' or 'increase q until modularity is above a certain threshold'.

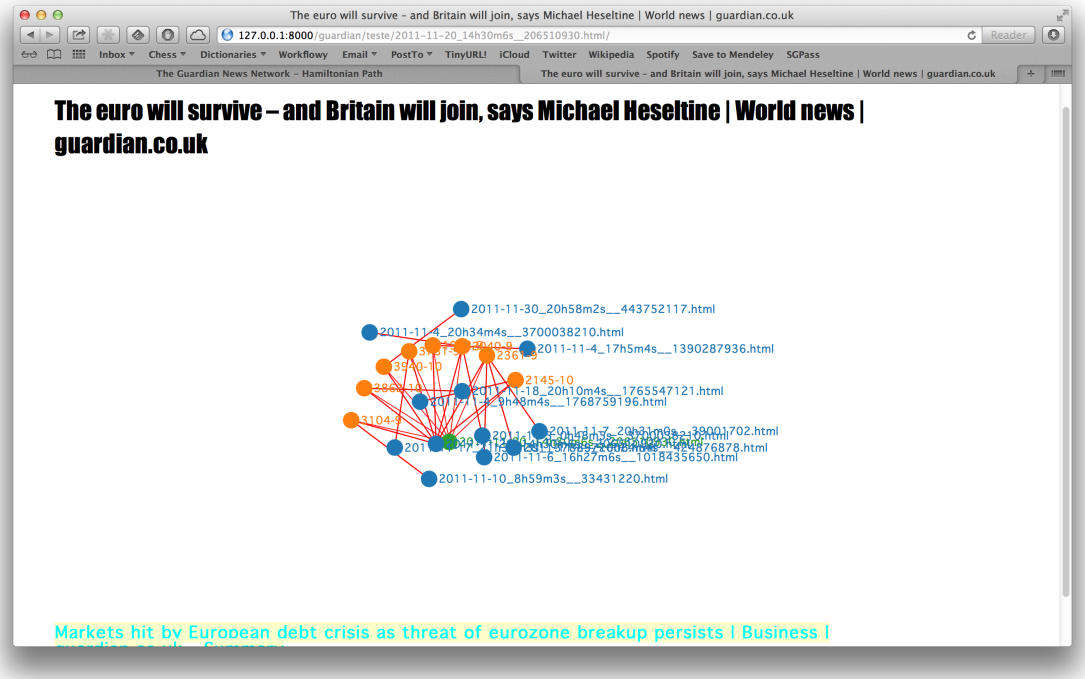


Figure 10.11: Q-analysis visualisation software displaying the active document (green) and the connected documents (blue) via their shared faces (orange)

Each of these cases depends clearly on the problem in hand and the user needs. In the

above case the resulting graphs for *q=8* and *q=9* are very sparse and from the discussion of the strategy a better solution is to choose *q=5* (full breakdown of the topics can be found in appendix B. A visualisation application was developed to allow the navigation through the resulting structure. This is illustrated in figure 10.11 where the connectivity between the active document (in green) is put in evidence through the shared faces (in orange) and the neighbour documents.

It is clear from the previous graphs and results that there's not a definitive recipe for the thresholds choosing. Different problems may present different characteristics, but in general we have shown that by using modularity optimisation techniques on top of a graph construction that takes into account the connectivity of the underlying system, it is possible to obtain high confidence. Manual inspection of the resulting topics, even just the titles of the news (as show in appendix B) clearly indicates that very few *noise news* are present. These noise news tend to be sparsely connected and usually only under very broad terms. A value of $q = 5$ seems to be a good starting point to filter most of the noise news, showing that news that have more common tags will be highly related. These common tags form the underlying structure of the topics found and allow a navigation through them in a coherent way.

Q-analysis gives a descriptive overview of the structure of the system, in terms of the local connectivity of the news stories. It does not try to separate them in clusters and on the other hand gives the researcher a framework on which to base a discourse on the 'traffic' that this structure supports.

By applying a clustering algorithm on top of this structure one adds value to the analysis by seeing if the structure that *Q*-analysis describes, is in reality a single structure, or if it is composed of modules that are somewhat analysable alone.

The two techniques combined in algorithm 3 produce a filtering solution for the problem of identifying relevant topics in online media production. The technique is robust enough to be used with different datasets of online textual documents, because the relation between document and labels can be obtained through a variety of techniques (human input, machine learning, latent Dirichlet analysis, etc.)

In Chapter 11 a novel approach is presented where one takes advantage of the structural connectivity of the simplicial complex formed by the news and the tags entered by the journalists. In that case one does not filter the lowly connected news but takes advantage of the eccentricity between nodes to define paths across all news.

Chapter 11

Hamiltonian paths in *Q*-analysis eccentricity matrices

This section introduces an algorithm developed in collaboration with Vitorino Ramos for the symmetric travelling salesman problem (TSP) based on Ant Colony Optimisation (ACO) Systems that was then applied to the news stories published by the newspaper **The Guardian**. This was done with the objective of finding a navigation system between news that would minimise the distance travelled in terms of Q-analysis eccentricity[19] [20]. The algorithm uses negative pheromones deposited on the paths that are sub-optimal allowing them to work as 'no-entry' markers and speeding up the traditional ant colony system (ACS) algorithm.

The main objective of this work is to introduce a novel algorithm for the construction of Hamiltonian paths of the news published by *The Guardian*. The algorithm presented here proved to excel, being better than traditional Ant Colony Systems in benchmark problems like the travelling salesman problem (TSP) and was applied to the construction of tours of news items.

The TSP problem and the ant based algorithms have developed a set of metaphors to translate the mathematical nomenclature into problem specific nomenclature. For the readability of the following sections it is necessary to clarify those here. While the TSP problem is the search of a Hamiltonian path in a undirected graph, it is usually put in terms of agents visiting cities. Ant based algorithms, use the metaphor of ants for these agents,

[19] This work has been presented at the ECCS'11 conference in Vienna (Rodrigues et al., 2011; Ramos et al., 2011) and at the HAIS'13 (Ramos et al., 2013) conference in Salamanca, and was published in Lecture Notes in Computer Science (Ramos et al., 2013)

[20] slides available at http://www.davidrodrigues.org/pdfs/2013/salamanca-hais2013-77-2ndOrder.pdf and a pdf copy is available at http://arxiv.org/abs/1306.3018

and pheromones for the weights of edges connecting two vertices (cities). This means that usually one refers to cities, ants and pheromones when in mathematical terms one would refer to vertices, agents and weights.

11.1 Second order ant colony system in TSP

Standard stigmergic approaches to swarm intelligence encompasse the use of a set of stochastic cooperating ant-like agents to find optimal solutions, using self-organised stigmergy as an indirect form of communication mediated by a singular artificial pheromone.

Agents deposit pheromone-signs on the edges of the problem-related graph which give rise to a family of successful algorithmic approaches entitled Ant Systems (AS), Ant Colony Systems (ACS), among others. These mainly rely on positive feedback, to search for an optimal solution in a large combinatorial space.

The present work shows how, using two different sets of pheromones, a second-order co-evolved compromise between positive and negative feedback achieves better results than single positive feedback systems.

This follows the route of very recent biological findings showing that forager ants, while laying attractive trail pheromones to guide nest mates to food, also gain foraging effectiveness by the use of pheromones that repel foragers from unrewarding routes.

The solution construction for the travelling salesman problem (TSP) can be accomplished through two different heuristics:

Firstly, there are *building heuristics*. These heuristics use a starting vertex and build, according to some pre-defined rule, a solution through the addition of successive vertices. This is the case of the nearest neighbour heuristic where at each step of the construction the nearest vertex of the previous city, that is not yet in the list, is appended.

Secondly, there are *optimising heuristics*, that aim to improve given solutions. In this class, for an initial given solution, the heuristic usually tries to improve it by replacement/mutation of the connections between the existing vertices. If the result of the operation improves the solution, then it is adopted, otherwise it is discarded. Examples of these heuristics include the *2-opt*, *3-opt*, and *Lin-Kerningham* algorithms, where the number of connections exchanged is two, three, or a variable number, respectively (Dorigo and Gambardella, 1996).

It has been verified experimentally that *building heuristics* usually perform worse

than *optimising heuristics* for these kinds of problems (Dorigo and Gambardella, 1996, chap. VI and references therein). In any case for *optimising heuristics* to work is always necessary to feed them with a pre-computed solution given by a *building heuristic* (even if it is a random one).

In the TSP case the *building heuristic* is performed by the ants and by the stigmergic variable (pheromone deposited on the cities connections). The solutions built by this process can than be fed into an *optimising heuristic*, such as *2-opt* to improve it. This combination of heuristics produces better results than each heuristic in isolation.

11.2 Formulation of the second order swarm intelligence (SOSI) algorithm

Traditional approaches to the TSP via Ant Systems include only a positive reinforcement pheromone. This new approach uses a second negative pheromone, that acts as a marker for forbidden paths. These paths are obtained from the worse tour of the ants and this pheromone then blocks access of ants in subsequent journeys. This blockade is not permanent and, as the pheromone evaporates, allows for paths to be tried again for better solutions. Also, during the local update heuristics where pheromone is deposited in the path after each ant step, this negative pheromone deposition promotes the selection of other alternatives by other visiting ants which acts as a mechanism to promote variability.

Modified state transition rule AS

$$p_k(r,s) = \begin{cases} \frac{[\tau^+(r,s)]^\alpha [\eta(r,s)]^\beta [\tau^-(r,s)]^{\alpha-1}}{\sum_{u \in J_k(r)} [\tau^+(r,u)]^\alpha [\eta(r,u)]^\beta [\tau^-(r,u)]^{\alpha-1}} & \text{if } s \in J_k(r) \\ 0 & \text{otherwise} \end{cases} \tag{11.1}$$

Modified State Transition Rule in ACS

$$p_k(r,s) = \begin{cases} \arg\max_{u \in J(k)} \{[\tau^+(r,u)]^\alpha [\eta(r,u)]^\beta [\tau^-(r,u)]^{\alpha-1}\} & q < q_0 \\ S & \text{otherwise} \end{cases} \tag{11.2}$$

Modified global updating rule

$$\tau^+(r,s) \leftarrow (1-\rho^+) \cdot \tau^+(r,s) + \rho^+ \Delta\tau^+(r,s) \tag{11.3}$$

$$\Delta\tau^{+}(r,s) = \begin{cases} L_{gb}^{-1} & \text{if } (r,s) \in \text{Global best tour} \\ 0 & \text{otherwise} \end{cases} \tag{11.4}$$

$$\tau^{-}(r,s) \leftarrow (1-\rho^{-}) \cdot \tau^{-}(r,s) + \rho^{-}\Delta\tau^{-}(r,s) \tag{11.5}$$

$$\Delta\tau^{-}(r,s) = \begin{cases} nL_{gb}^{-1} & \text{if } (r,s) \in \text{Global worse tour} \\ 0 & \text{otherwise} \end{cases} \tag{11.6}$$

Modified local updating rule

$$\tau^{+}(r,s) \leftarrow (1-\rho) \cdot \tau^{+}(r,s) + \rho\Delta\tau^{+}(r,s) \tag{11.7}$$

$$\tau^{-}(r,s) \leftarrow (1-\rho) \cdot \tau^{-}(r,s) + \rho\Delta\tau^{-}(r,s) \tag{11.8}$$

11.3 Outline of the algorithm

Algorithm 4 Pseudocode for the 2^{nd} order ant colony optimisation heuristic (SOSI)

Require: max_iter, $tsp\_problem$, num_ants, $num\_vertices$, and other parameters
  Initialise parameters
  Initialise both positive and negative pheromone distributions
  $iter \leftarrow 1$
  while $iter \leq max\_iter$ **do**
    Position each ant on a randomly chosen starting vertex
    for $step = 1 \rightarrow num\_vertices$ **do**
      for $ant = 1 \rightarrow num\_ants$ **do**
        if $q < q_0$ **then**
          Apply exploitation strategy (eq. 11.2)
        else
          Apply exploration strategy (eq. 11.1)
        end if
      end for
      Local pheromone update (eqs. 11.7 and 11.8)
    end for
    if $localSearch == True$ **then**
      Perform local search heuristic on solutions found
    end if
    Calculate ants best and worse solutions lengths
    Update global positive pheromone (eqs. 11.3 and 11.4)
    Update global negative pheromone (eqs. 11.5 and 11.6)
    $iter \leftarrow iter + 1$
  end while

11.4 Validation of the SOSI algorithm on benchmark problems

The new algorithm was exhaustively tested on a series of well-known benchmarks over hard NP-complete Combinatorial Optimisation Problems (COP's), running on symmetrical Travelling Salesman Problems (TSP). Different network topologies and stress tests were conducted over low-size TSP's (*eil51.tsp; eil78.tsp; kroA100.tsp*), medium-size (*d198.tsp; lin318.tsp; pcb442.tsp; att532.tsp; rat783.tsp*) as well as large sized ones (*fl1577.tsp; d2103.tsp*) [numbers in the previous well known tsp problems refer to the number of nodes in each problem].

It is clear from table 11.1 that the 2^{nd} order AS performs at least equally, if not better,

Table 11.1: Test bed and optimal results for the TSP problem

problem	**n.º of nodes**	**standard ACS**	2^{nd} **order**$^{+}$ AS	**optimal tour**
eil51.tsp	51	427.96	428.87	426
eil78.tsp	78	**	544.34	538
kroA100.tsp	100	21285.44	**21285.44**	21282
d198.tsp	198	16054	**15900.20**	15780
lin318.tsp	318	42029***	42683.90	42029
pcb442.tsp	442	51690	**51464.48**	50778
rat783.tsp	783	9066	**8910.48**	8806
fl1577.tsp	1577	23163	**22518**	22249
d2103.tsp	2103	-	81151.9	80450

Optimal tours from http://comopt.ifi.uni-heidelberg.de/software/TSPLIB95/STSP.html
+ Average over 20 runs and limited to 1000 iterations
** Value for similar problem eil75,.tsp - 542.37
*** uses 3-opt local search

than the standard ACS. It is clearly seen that the averages of the runs (bold) are better than the traditional ACS.

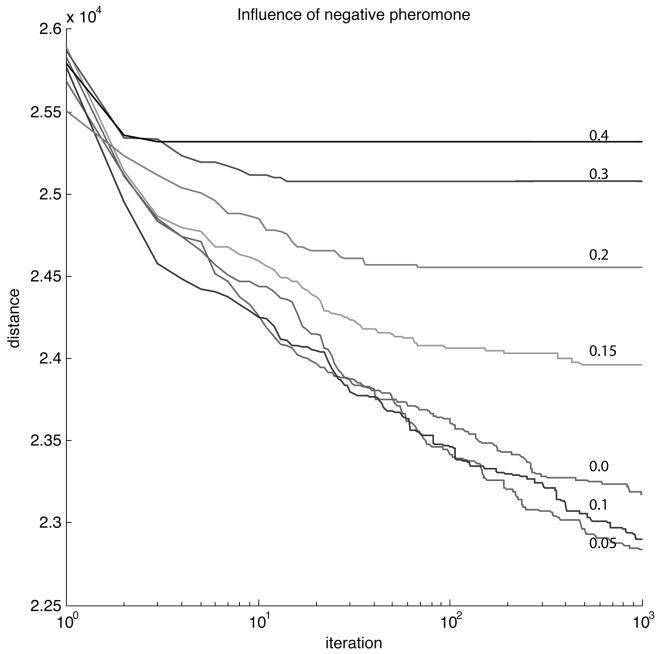


Figure 11.1: Influence of negative pheromone on *kroA100.tsp* problem (values over the lines represent $1.0 - \alpha$)

The evolution of different ratios of negative pheromone to positive pheromone was investigated and it was verified that a small amount of negative pheromone applied as a 'no-entry' signal indeed produces better results. However this positive effect is cancelled if

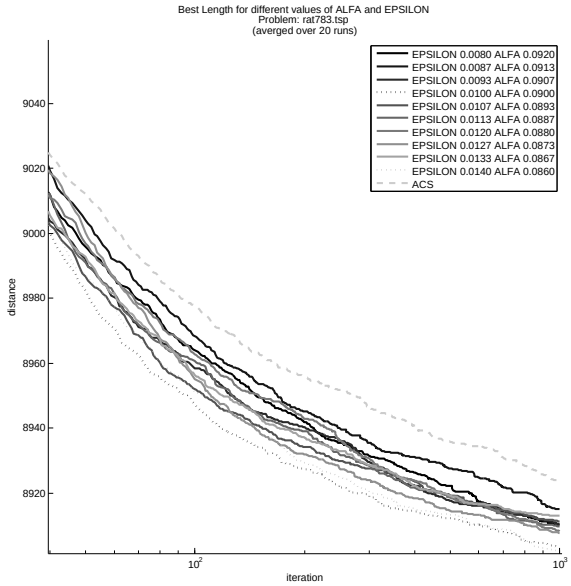


Figure 11.2: Influence of negative pheromone on *rat783.tsp* problem

the ratio of negative pheromone is high when compare to the positive pheromone.

The effect of negative pheromone can be observed both in figure 11.1 and figure 11.2 where one can see that small amounts of negative pheromone produce better results and quicker convergence to those results. On the other hand if one increases the ratio of negative pheromone to higher values then it is not possible to reap the benefits of the 'no-entry' signal and the system performs less well.

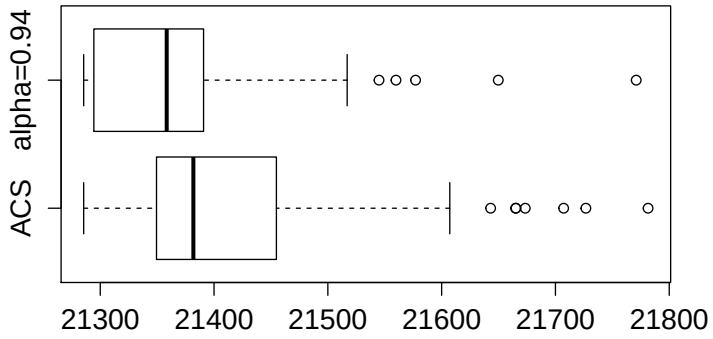


Figure 11.3: Boxplot of 120 runs comparing the influence of negative pheromone on *kroA100.tsp* problem with traditional ACS

The detailed analysis of the *kroA100.tsp* problem showed that the effect of the negative pheromone is statistically significant. Comparing 120 runs with $\alpha = 1.0$ (equivalent to a traditional ACS and $\alpha = 0.94$, a *p*-value of 3×10^{-4} was obtained. This result is summarised in figure 11.3, where the traditional ACS with the 2^{nd} order approach are compared.

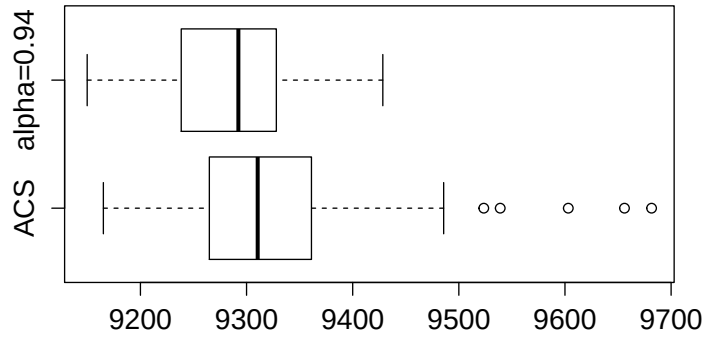


Figure 11.4: Boxplot of 70 runs comparing the influence of negative pheromone on *rat783.tsp* problem with traditional ACS

The same results where observed for problem *rat783.tsp* when comparing 70 runs of the ACS ($\alpha = 1.0$) with 70 runs of the 2nd order approach (with $\alpha = 0.94$) in figure 11.4. The two samples means were tested for statistical significance resulting in a *p*-value of 2.2×10^{-3}.

Both these examples show that on average the 2^{nd} order approach performs better than traditional ACS. This effect of the negative pheromone is important but cannot be extended further as to dominate the solving strategy, making results worse. This can be seen clearly on figure 11.5 where further diminishing of alpha (giving more weight to negative pheromone as a consequence) produces worse results.

The algorithm was also tested in dynamical stress tests, where at predefined intervals a certain amount of cities is changed in order to see how the system reacts to the change and recovers to optimal solutions. This was done for the problem *fl1577.tsp*. Figure 11.6 shows that the new algorithm is quick to recover from those perturbations. The randomisation of the cities was done every 250 iterations. Also different percentages of cities ranging from 20% to 100% were changed in those steps. The figure clearly shows that the algorithm is

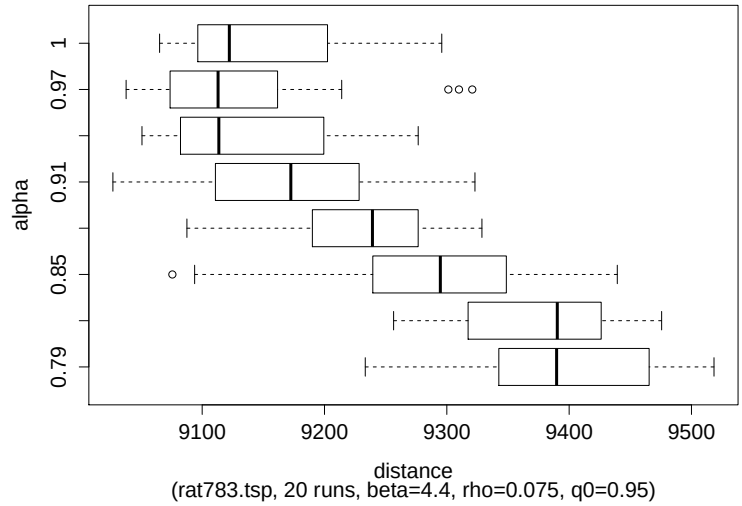


Figure 11.5: Influence of negative pheromone $(1 - \alpha)$ on the TSP problem *rat783.tsp*

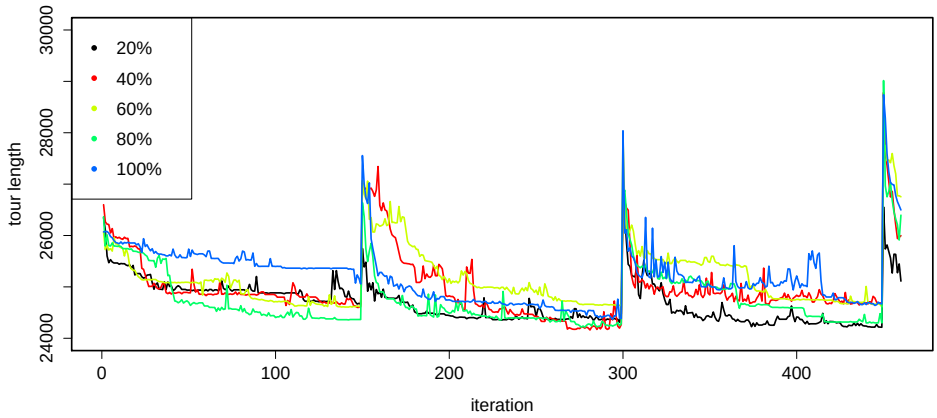


Figure 11.6: Recovery times of the Dynamical stress tests over fl1577.tsp problem (1577 nodes) - 460 iterations - Swift changes at every 150 iterations (20%=314 nodes, 40%=630 nodes, 60%=946 nodes, 80%=1260 nodes, 100%=1576 nodes)

capable of recovering quickly from these perturbations, even for full randomisation of the problem.

11.5 Application of the SOSI algorithm to the news

11.5.1 Eccentricity as a distance measure between simplicies

In the analysis of the timeline of The Guardian (Chapter 9) the system used feature vectors based on frequency of words and them computed similarity between documents based on those feature vectors. This is a purely statistical approach that requires great computational power and that is difficult for problems that have large feature vectors and many documents. Feature vectors with 100,000 or more items are common and computing similarities between these documents becomes cumbersome. Instead of computing distance (or similarity) matrices between documents from feature vectors, this approach explores the possibility of inferring the distance between documents from the Q-analysis description done in section 10. Q-analysis is a very natural notion of connectivity between the simplicies of the structure and in the relation studied, documents are connected to each other through shared sets of tags entered by the journalists. Also in this framework, eccentricity is defined as a measure of the relatedness of one simplex in relation to another.

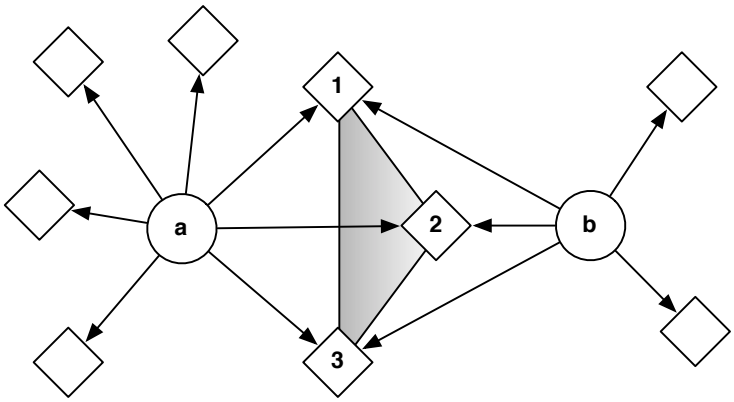


Figure 11.7: Two simplicies a and b connected by the 2-dimensional face, the triangle $\{1, 2, 3\}$.

More formally, Q-analysis defines the eccentricity, $ecc_{a,b}$, between 2 simplicies a and b as:

$$ecc_{a,b} = \frac{|a| - |a \cap b|}{|a|} \tag{11.9}$$

were $|a|$ means cardinality of the set a.

It is clear from equation 11.9 that eccentricity of a simplex is defined in relation to the other simplex and therefore $ecc_{a,b}$ is not equal to the $ecc_{b,a}$ except when $|a| = |b|$.

As an illustration consider the simplex (circles) a and b from figure 11.7. In this case they are connected by the bi-dimensional face made of the shared vertices $\{1, 2, 3\}$. The value of $|a| = 7$, $|b| = 5$ and $|a \cap b| = 3$. The eccentricities calculated by equation 11.9 are $ecc_{a,b} = 4/7$ and $ecc_{b,a} = 2/5$ showing that eccentricity is not symmetrical and is a measure of how distant one simplex is from being totally contained in another simplex.

Eccentricity thus defined can be thought of as a directed distance measure between simplicies. This measure arises from the data structures and is therefore highly representative of the structural properties of the system.

By using the eccentricity matrix as the distance matrix the problem of computing statistical features of the documents is avoided. This was done previously in section 9 where feature vectors were used to compute a clustered timeline of events.

11.5.2 Application to *The Guardian* news

The application of the algorithm developed in the previous sections to the news published by The Guardian was done on a monthly basis and the Hamiltonian paths were constructed using the SOSI algorithm to form a navigation structure over the simplicial complex of the accumulated news. This one dimensional mapping of the news was achieved by using as the distance matrix for the algorithm the eccentricity matrix. Although in this context there is no natural relation between the data points and the euclidian 2D space usually used in the ACO metaphor, the eccentricity matrix reflects the fraction of vertices that would need to either drop or connect to the other simplex in order to achieve low eccentricity (a simplex has zero eccentricity with itself $ecc_{a,a} = 0$).

A software web application (illustrated in figure 11.8) was developed to allow the convenient navigation through the Hamiltonian path and is part of the Theseus package. In appendix C the Hamiltonian path for the month of November 2011 is given as an illustration. The applications uses the best tours from the second order swarm intelligence algorithm to display news to the user in a sequential manner allowing for a navigation through highly related (low eccentricity) news items.

A detail of the November 2011 analysis is visible in figure 11.9 where each news title is interleaved by a line with tags. The tag line shows the tags that are present in the first story. Underlined are the tags that are also present in the second story. In this way it is

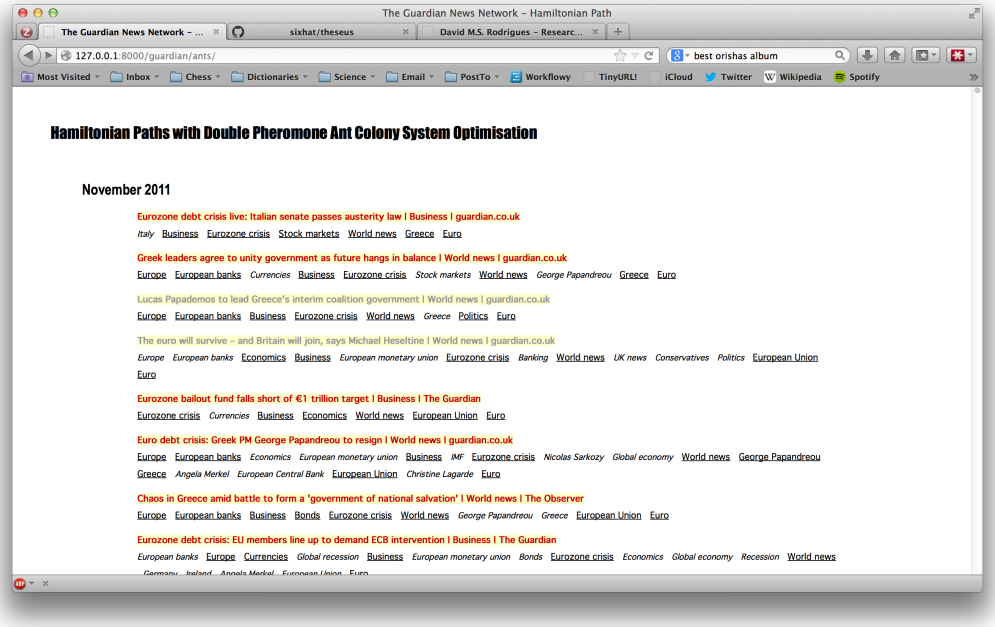


Figure 11.8: Application developed to visualise the Hamiltonian paths found by the second order algorithm

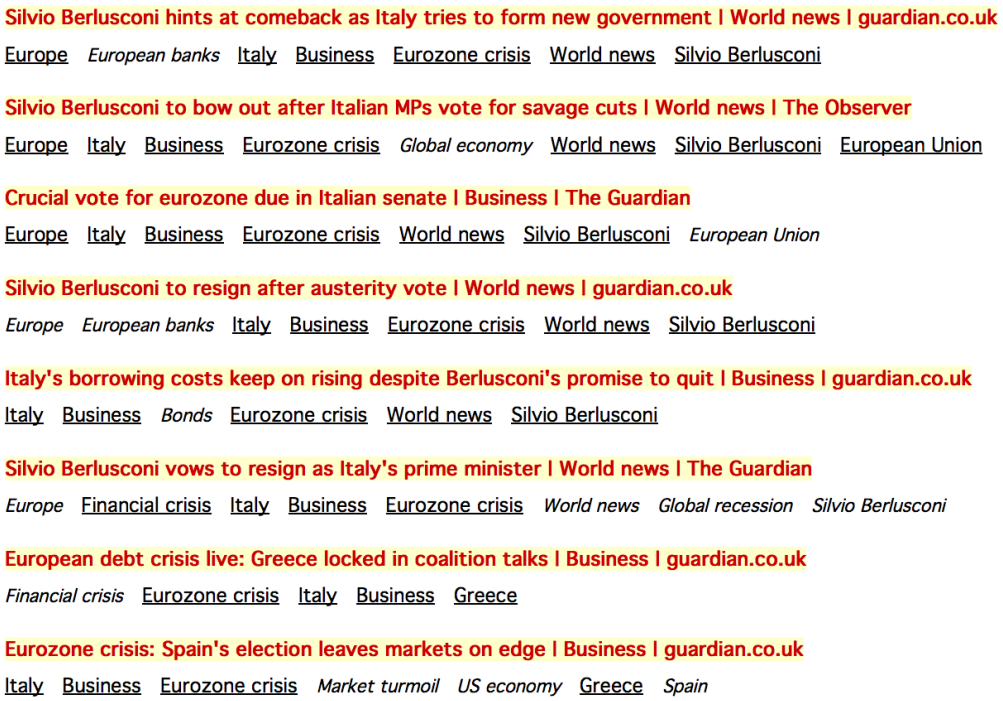


Figure 11.9: Details of the application developed to find the Hamiltonian paths with the 2^{nd} order swarm intelligence algorithm

easy to see the eccentricity of the first story in relation the the second in this sequence. For example the eccentricity between the first story in relation to the second story in figure 11.9 is $1/7$ because only the tag *Europan banks* from the first story is not present in the second story while the remaining six are. This helps the user understand why these stories are related by showing the degree of connectivity between the stories. In the developed application, the title of the news is an hyperlink that allows the user to navigate to the original guardian news story.

11.6 Concluding remarks

It is shown that the new co-evolved stigmergic algorithm compared favourably against the benchmark. The inclusion of a negative pheromone acting as a 'no-entry' signal in the strategy of construction of solutions is beneficial as the convergence to optimal solutions is quicker, as shown in figure 11.2, while achieving better results (figures 11.3 and 11.4). The algorithm was able to equal or greatly improve every instance of those standard algorithms.

The new algorithm comprises a second order approach to Swarm Intelligence, as pheromone-based no entry-signals cues were introduced, coevolving with the standard pheromone distributions (collective cognitive maps (Ramos et al., 2006)) in the aforementioned known algorithms.

The use of the negative pheromone is limited to small quantities (α close to 1, but not 1, in which case it would become a pure ACS) and cannot be extended to a point of dominance in the search strategy as shown in figure 11.5. The results found for the TSP problems in that case are severely worse. This implies that the use of a negative pheromone strategy has to be fine tuned so as not to dominate the search strategy. This is done with the introduction of the parameter α that balances the weight of the two pheromones deposition in equations 11.1 and 11.2.

The new algorithm also performed well in stress tests where perturbations in the problem fed to the algorithm where made at certain points in the experiments. The algorithm quickly recovered from those perturbation as can be seen in figure 11.6 for varying intensities in the perturbation. This indicates good prospects for the future success of this algorithm in dynamical problems where the nature of the problem is constantly evolving.

The application of the algorithm to the Hamiltonian path finding problem in news

published online reveals natural sequences of related stories as shown by the examples in figures 11.8 and 11.9. The choice of eccentricity as a measure of distance between two documents is a natural choice as it derives from the structural connectivity of the simplicial complexes obtained in Q-analysis. It also has the advantage that it is easily calculated and does not require the computation of statistical transformation as in the case of the feature vectors used in the approach of the timeline construction.

This work has implications in the way large combinatorial problems are addressed as the double feedback mechanism shows improvements over the single-positive feedback mechanisms in terms of convergence speed and of major results.

Chapter 12

Conclusions

News published online by both traditional and new publishing companies are changing how the world is perceived in unprecedented ways. The massive availability of information makes extraction and summarisation a daunting task for any automated computer system. Because of this, there is a need for new ways to find, explore and consume information. To understand these new ways more knowledge about the structural properties of this emerging system is needed. In this research it is shown that understanding how these news corpora structure themselves helps in designing better computer aided algorithms to find, cluster, read and digest the news. It has been shown how different approaches can help understand the structural properties of modern news publishing. Through a series of case study experiments several applications were developed that reveal insights of the complex systems of news publishing that will benefit future applications acting on similar corpora.

12.1 Main Results

Part II of this document shows how four different hybrid connectivity based approaches were developed and used to analyse the news produced by online newspapers. The summary of contributions is explained in the following paragraphs and summarised in table 12.1.

In Chapter 8 it was shown that a mutual information based measure (Meilă's variation of information), coupled to a model where each node is added dynamically to an adaptive network, allows the identification of topics that emerge and die over time in the Portuguese newspaper Público. This technique has been shown to be useful in identifying clusters from datasets *a posteriori*, as the variation of information reacts strongly to big changes in the structure of the network. This occurs when a large cluster of news items disappears causing a cascade in the variation of information between consecutive states of the network.

In any case the destruction of a cluster of news items corresponds to the disappearence of a topic of highly related news stories. It was shown that according to the model, the topics of the daily newspaper **Público** have an approximate lifetime of 17h reflecting the daily character of the newspaper.

The work presented in Chapter 9 on **The Guardian**, showed how the stories published online can be clustered into timelines according to their similarity. This used a k-means algorithm to cluster news items that are processed as feature vectors from the frequency of words (more precisely the 'term frequency inverse document frequency' measure **tf.idf**).

In Chapter 10, it was shown that by using the intrinsic connectivity of the news stories published online one is able to reveal the structure behind novel topic clusters. By applying a Q-analysis to the system of news published by the newspaper it was demonstrated that the algorithm can filter sparsly connected stories, i.e. 'noise news' and get a resulting structure that is highly modular. Then, the application of a clustering technique based on modularity optimisation reveals the topics, and the clustering results are free of noise elements. An example of the resulting clustering is presented in Appendix B.

In Chapter 11, a Second Order Swarm Intelligence approach to the symmetrical TSP problem was presented, which described a new idea for a second order feedback mechanism based on the inclusion of negative pheromones. While traditional ACS algorithm rely on a double feedback mechanism with only one kind of pheromone, the proposed approach uses a second negative pheromone based on recent findings on real ants. This other pheromone acts as a 'no-entry' signal and permits fast and deeper exploration of the search space of the the TSP problem. This new algorithm was benchmarked against known NP-complete combinatorial Optimisation Problems, running on symmetrical TSP and the proposed technique compares favourably against these benchmarks. Also, the algorithm is quick to respond to perturbations in the problem at hand. Stress tests demonstrated that the stigmergy approach is useful for application to dynamical changing problems.

The results obtained with the development of this new algorithm where applied in Chapter 11 to the construction of Hamiltonian paths through the news published online. The discovery of the Hamiltonian paths uses the eccentricity between the simplices as a measure of the distance between documents. It was shown that this approach, using the natural connectedness of the system, reveals a very good coherence in the navigation of documents, providing an alternative view to a time series based navigation of the news. The usage of the eccentricity as the local measure of distance avoids the computationally

expensive usage of feature vectors as in chapter 9 or the arbitrary definition of a similarity measure like the one used in chapter 8 for the variation of information and adaptive network case.

12.2 Main contributions, advantages and disadvantages

The algorithms presented in this thesis have both advantages and disadvantages depending on the application. Table 12.1 presents a summary of the hybrid methods developed in this work and their respective advantages and disadvantages.

Table 12.1: Advantages and disadvantages of the proposed hybrid connectivity based approaches

Methodology	Advantages	Disadvantages
Adaptive Networks and Variation of Information	Allows different similarity measures to be used. Easy to parameterise through TTL.	Analysis a posteriori as only when clusters disappear can one see high VI. Connectivity not directly extracted from data but from a similarity function
Feature Vectors and Timelines with k-means	Connectivity given directly from word frequency. k-means easy to implement	k-means method is non-deterministic. Number of clusters needed to be known *a priori*. Feature vectors usually very large with more than 10,000 entries.
Q-analysis and Modularity Optimisation	Connectivity based on structural properties of the bipartite graph of the data. Applicable to many clustering methods.	Filtering noise news requires human definition of threshold q. Modularity has resolution limits.
Q-analysis and Second Order Swarm Intelligence	New algorithm with potential to explore the negative pheromone idea in future work. First results proved interesting. SOSI algorithm can be applied to dynamical problems as solutions are found dynamically. Uses Eccentricity as a direct measure of distance between news. No data manipulation, keeping representation close to original data.	Computationally more expensive than traditional ACS. Eccentricity of documents has to be recomputed as each new news story is added and is not a metric space.

The contributions presented in this work show that by taking into account the connectivity patterns of the complex systems studied, one obtains better insights to the topics of the news stories being published online. In particular, the methods based on Q-analysis take advantage of the high-dimensional representation of the relations between the elements of the system to produce relevant results.

Beyond the contributions listed above, the contribution in terms of software is also relevant. For the execution of this project it was necessary to develop tools to gather, store, analyse and display the news collected. This software named *Theseus* was released in an open source repository[21] and is briefly described in Appendix A.

The work developed on the Second Order Swarm Intelligence (SOSI) in Chapter 11 shows how a stigmergy based algorithm can be applied to the exploration of text based documents. From this analysis it became obvious that the search for Hamiltonian paths in the corpus of documents reveals useful relations, mainly when using the algorithm with a measure of distance based on the eccentricity between documents. The SOSI algorithm is a graph traversal algorithm and not a clustering one as some techniques of clustering noted in section 6.3, but the idea of the double feedback mechanism can in the future be applied to existing clustering algorithms without the need to develop a new second order framework from scratch. In this case the graph traversal algorithm was able to produce significant sequences of related news providing a novel way to navigate through online news.

The SOSI algorithm is also very useful as an extension of the traditional Ant Colony Systems algorithm, as it proved to be faster and gives better results than the traditional algorithm. This opens the door for future development in applying the bio-inspired double pheromone techniques.

The main limitation of the Variation of Information method used in Chapter 8 is that it only signals the structural change in the clustering at the moment that the cluster disappears from the adaptive network which is being dynamically built. For a large cluster to disappear (High VI) it is necessary that all its members drop their Time To Live (TTL) to zero. This means that the detection of topics by the Variation of Information method alone carries a delay in time that might not be useful for real time applications. Also, the method requires the definition of a similarity function between two different stories by the user. This again puts the burden of defining when two stories are similar on manual human intervention. An alternative to the manual definition of distance is to explore the high dimensional aspect of the documents as was done for The Guardian newspaper. This case study is still useful nonetheless as it provides a quick and dynamical network approach to news evolution over time. The model depends also on the chosen parameter for the TTL of the new nodes. This is equal for all types of documents and a model with topic specific TTLs would probably

[21] Public repository of Theseus is available at https://github.com/sixhat/theseus

improve the results as intuition suggests that sports news has a different TTL from news about economy. This model of dynamic networks would benefit future research.

The same problem is observed in the timeline method used for the analysis of **The Guardian** in Chapter 9. In this case, the *k*-means method creates the additional difficulty of forcing the user to predefine the number of clusters. This is not always possible and other strategies were pursued to try to circumvent these limitations, namely by using clustering techniques that do not require *a priori* knowledge of the number of classes in the system. The user input of the number of clusters forces the system into situations where news are lumped into the same cluster and the feature vector approach based on bags of words does not provide the right dimensional representation for a good extraction of the modules of the resulting structure. Future work on the *k*-means algorithm could take into account the high dimensional representation of the simplicial complex generated from the bipartite relation between documents and tags.

Although the filtering and clustering system, developed in Chapter 10, does not pose the problem of pre-knowing the number of classes present in the system, the analysis showed that the cut-off value of q is very relevant. The connectivity of the news present in the newspaper is sparse and even for low values of q, many news items are removed from the system. This transition has to be analysed on a case by case scenario which poses a limitation to a fully automated system. However automatic ICT systems can define a threshold value for the maximum number of news items removed by the algorithm: e.g. increase q until 5% of news are removed. More sophisticated strategies could find the abrupt transition of the system as was observed in Figure 10.2 and subsequent figures. The analysis shows consistent behaviour through the 3 years of the dataset, meaning that probably human intervention would be needed very rarely and that after initial tuning of the system little intervention would be required. Another limitation of the developed system is that it uses modularity optimisation to do the clustering. This technique has some known problems described in 5.2. Future work could use other mechanisms of identifying clusters based on other ideas. Finally, eventually the second order framework developed for finding Hamiltonian paths could be adapted to develop new methods for clustering .

The Theseus software was developed to answer precisely the questions posed in this thesis and therefore has some limitations on what it can do and on its maturity for general purpose usage. The software is usable, but requires proficiency in linux/unix environments. Also, documentation for the present software is not fully written meaning

that interested users need to be able to read the original code provided at the Theseus repository and in the examples folder provided with the software. It was decided to develop Theseus in Python for ease of prototyping the models and experiments. Also, it would be possible to use a vast collection of existing open source libraries and integrate them in the workflow of the scientific research. This is easily noticed in the Theseus dependencies section A.3.3 in page 146. Also, the Python language implements natively the notion of sets, and many mathematical operations usually needed in set theory are available in Python. This was as strong argument for choosing Python as the development language for Theseus. Readability of the code, and rapid prototyping of the developed algorithms where decisive considerations over code optimisation procedures, parallelisation, and distributed computing techniques. It is therefore relevant that future research allocates funds and resources to the improvement, maturation and dissemination of this software. The licensing of this software under an Open Source license might help the formation of a community of contributors and users that will help in its future development.

12.3 Perspectives

This research set out to investigate the structural properties of news published in online newspapers, namely through the natural connectivity presented in the system being studied. The hybrid connectivity based approaches presented in this thesis provide a natural guide to future research on the same topics.

The approach of using adaptive networks with variation of information should be explored further. Although the technique only allows for *a posteriori* analysis of the news, it is potentially very interesting as it allows for the exploration of information-based measures in the publication of the news. These can act as *signal detectors* and research on how this information can be used to move from *a posteriori* detection to real-time detection has great potential in the application to breaking news detection or finding scientific novelty.

In the field of bioinspired algorithms it is my belief that many opportunities arise by mimicking nature to develop new algorithms based on the triad exploitation, exploration and evolution. The work presented here is an example of an artificial system where exploitation and exploration are used to produce a novel algorithm. In future work, it will be of interest to study the role of evolution in the development of these algorithms and

to find new biologically observed behaviours that can be mapped into machine learning techniques to improve the quality of the solutions and explanations of the multilevel systems one studies.

Structural properties of complex systems are of upmost importance and understanding how the structural properties affect the dynamics that these systems exhibit is certainly a future research topic that will be thoroughly investigated. Q-analaysis is a sound descriptive language of the existing structure but future research should mainly focus on hypernetwork theory as it spins from Atkin's hypotheses that the topology of the backcloth constrains the dynamics of the traffic. It will be of great interest to test the hypernetwork theory as the modelling process of larger multilevel systems.

Future research in this field will always need to deal with large datasets, which are often labelled *big data*. Until now, most analysis on *big data* focused on the statistical properties of the data, and it is considered that right now it is the time of *data rich* versus *theory rich*. Future work should try to reconcile these two notions and develop new theories and models of *big data*. This would be of great interest in dealing with the amount of data that is being collected from the Internet. It is my belief that multilevel systems and a theory for automated extraction of the intermediate levels in multilevel systems that is computationally tractable is of upmost importance and will in the future be of great importance.

Hybrid approaches combine the best of different techniques to solve problems. It is my belief that in the future more and more hybrid approaches will emerge to tackle the high dimensionality of the problems faced by science. As these techniques can come from many different academic fields, it is important that interdisciplinary studies are able to have sufficient know how about them. Future research on the applicability domains of these hybrid approaches will benefit the methodologies used in complexity science and certainly be the basis for innovative discoveries about the studied problems.

Bibliography

Abraham, A. and Ramos, V. (2003). Web usage mining using artificial ant colony clustering and linear genetic programming. In *Proceedings of the Congress on Evolutionary Computation*, pages 1384–1391, Australia. IEEE Press. 67

Abul Hasan, M. J. and Ramakrishnan, S. (2011). A survey: hybrid evolutionary algorithms for cluster analysis. *Artif. Intell. Rev.*, 36(3):179–204. 66

Aggarwal, C. and Zhai, C. (2012). A survey of text clustering algorithms. In Aggarwal, C. C. and Zhai, C., editors, *Mining Text Data*, pages 77–128. Springer US. 19, 22

Al-Kabi, M., Halalsheh, N. Z., Dabour, M., and Wahsheh, H. A. (2012). Arabic news: topic and novelty detection. In *Proceedings of the 3rd International Conference on Information and Communication Systems*, ICICS '12, pages 7:1–7:5, New York, NY, USA. ACM. 21

Albert, R. and Barabási, A. (2002). Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47–97. 24, 32

Albert, R., Jeong, H., and Barabasi, A. (1999). Internet: Diameter of the World-Wide web. *Nature*, 401(6749):130–131. 29

Albert, R., Jeong, H., and Barabási, A. L. (2000). Error and attack tolerance of Complex Networks. *Nature*, 406:378–382. 30, 32

Albert, R., Jeong, H., and Barabási, A. L. (2001). Errata: Error and attack tolerance of Complex Networks. *Nature*, 409:542+. 30

Allan, J. (2002). Introduction to topic detection and tracking. In Allan, J., editor, *Topic detection and tracking*, chapter Introduction to topic detection and tracking, pages 1–16. Kluwer Academic Publishers, Norwell, MA, USA. 20

Allen, R. B., Japzon, A., Achananuparp, P., and Lee, K. J. (2007). A framework for text processing and supporting access to collections of digitized historical newspapers. In *Proceedings of the 2007 conference on Human interface: Part II*, pages 235–244, Berlin, Heidelberg. Springer-Verlag. 20

Arenas, A., Díaz-Guilera, A., and Pérez-Vicente, C. J. (2006). Synchronization processes in complex networks. *Physica D: Nonlinear Phenomena*, 224(1–2):27 – 34. Dynamics on Complex Networks and Applications. 60

Atkin, R. H. (1972). From cohomology in physics to q-connectivity in social science. *International Journal of Man-Machine Studies*, 4(2):139 – 167. 3, 7, 37, 38

Atkin, R. H. (1974). *Mathematical Structure in Human Affairs*. Heinemann Educational Publishers, 48 Charles Street, London, 1 edition. 2, 37, 38, 42

Atkin, R. H. (1981). *Multidimensional Man*. Penguin Books. 2, 43

Atkin, R. H., Bray, R., and Cook, I. (1968). A mathematical approach towards a social science. *Essex University Review*, (2):6–8. 3, 7, 37, 75

Atkin, R. H., Johnson, J., and Mancini, V. (1971). An analysis of urban structure using concepts of algebraic topology. *Urban Studies*, 8(3):221–242. 7, 37, 44, 76

Baeza-Yates, R. A. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA. 59

Barabási, A. L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286:509–512. 2, 30

Beaumont, J. R. and Gatrell, A. C. (1982). *An introduction to Q-analysis*. Geo Abstracts, Norwich Norfolk. 3, 42, 43

Beeferman, D., Berger, A., and Lafferty, J. (1999). Statistical models for text segmentation. In *Machine Learning*, pages 177–210. 20

Beil, F., Ester, M., and Xu, X. (2002). Frequent term-based text clustering. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pages 436–442, New York, NY, USA. ACM. 19

Bjell, J., Canright, G., Engø-Monsen, K., and P. Remple, V. (2009). Topographic spreading analysis of an empirical sex workers' network. In Ganguly, N., Deutsch, A., and Mukherjee, A., editors, *Dynamics On and Of Complex Networks*, Modeling and Simulation in Science, Engineering and Technology, pages 97–116. Birkhäuser Boston. 32

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008. 56

Boettcher, S. and Percus, A. G. (2001). Optimization with extremal dynamics. *Phys. Rev. Lett.*, 86:5211–5214. 57

Bonabeau, E., Theraulaz, G., and Dorigo, M. (1999). *Swarm Intelligence: From Natural to Artificial Systems*. Santa Fe Insitute Studies In The Sciences of Complexity. Oxford University Press, USA, 198 Madison Avenue, New York, 1 edition. 61

Brandes, U., Delling, D., Gaertler, M., Görke, R., Hoefer, M., Nikoloski, Z., and Wagner, D. (2006). On Modularity – NP-Completeness and Beyond. Fak. f. Informatik (ITI) , Karlsruhe. 55

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30:107–117. 19

Broadbent, S. R. and Hammersley, J. M. (1957). Percolation processes. *Mathematical Proceedings of the Cambridge Philosophical Society*, 53(03):629–641. 31

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. (2000). Graph structure in the web. *Comput. Netw.*, 33(1-6):309–320. 29

Cachopo, A. M. D. J. C. (2007). *Improving Methods for Single-label Text Categorization*. PhD thesis, Instituto Superior Técnico - Universidade Técncia de Lisboa, Lisboa, Portugal. 79

Caldarelli, G. (2007). *Scale-free networks: complex webs in nature and technology*. Oxford University Press. 24, 30

Caldarelli, G. and Garlaschelli, D. (2009). Self-Organization and Complex Networks. In Gross, T. and Sayama, H., editors, *Adaptive Networks*, volume 51 of *Understanding Complex Systems*, pages 107–135. Springer Berlin / Heidelberg. 7, 35

Cardoso-Cachopo, A. and Oliveira, A. L. (2003). An empirical comparison of text categorization methods. In Nascimento, M. A., Moura, E. S. D., and Oliveira, A. L., editors, *String Processing and Information Retrieval*, pages 183–196. Springer Verlag, Heidelberg, DE. 17, 18

Cheng, J., Zhou, J., and Qiu, S. (2012). Fine-grained topic detection in news search results. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, SAC '12, pages 912–917, New York, NY, USA. ACM. 21

Chin, C., Duckstein, L., and Wymore, M. L. (1991). Factory automation project selection using multicriterion q-analysis. *Applied Mathematics and Computation*, 46(2):107–126. 48

Clauset, A., Newman, M., and Moore, C. (2004). Finding community structure in very large networks. *Phys. Rev. E*, 70:066111. Phys. Rev. E 70, 066111 (2004). xiii, xiv, 55, 56, 90, 91, 94, 98

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297. 18

Criado, R., Flores, J., García del Amo, A., Gómez-Gardeñes, J., and Romance, M. (2012). A mathematical model for networks with structures in the mesoscale. *International Journal of Computer Mathematics*, 89(3):291–309. 2

Criado, R., Romance, M., and Vela-Pérez, M. (2010). Hyperstructures, a new approach to complex systems. *International Journal of Bifurcation and Chaos*, 20(03):877–883. 2

Croes, G. A. (1958). A method for solving traveling-salesman problems. *Operations Research*, 6(6):791–812. 62

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, Complex Systems:1695. 90

Danon, L., Díaz-Guilera, A., and Arenas, A. (2006). The effect of size heterogeneity on community identification in complex networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2006(11):P11010. 56

Derenyi, I., Palla, G., and Vicsek, T. (2005). Clique percolation in random networks. *Physical Review Letters*, 94:160202. 51

Dharanipragada, S., Franz, M., Mccarley, J. S., Roukos, S., and Ward, T. (1999). Story segmentation and topic detection in the broadcast news domain. In *in Proceedings of the DARPA Broadcast News Workshop*. 20

Díaz-Guilera, A. (2013). European conference on complex systems 2013. In *European Conference on Complex Systems 2013 Book of Abstracts*. Complex Systems Society and complexitat.cat. 2

Diestel, R. (2005). *Graph Theory*. Springer, 3rd edition. 24

Dorigo, M., Bonabeau, E., and Theraulaz, G. (2000). Ant algorithms and stigmergy. *Future Gener. Comput. Syst.*, 16(9):851–871. 66

Dorigo, M. and Caro, G. D. (1999). The ant colony optimization meta-heuristic. In Come, D., Dorigo, M., and Glover, F., editors, *New Ideas in Optimization*, pages 11–32. McGraw-Hill, London, UK. 62

Dorigo, M., Di Caro, G., and Gambardella, L. M. (1999). Ant algorithms for discrete optimization. *Artif. Life*, 5(2):137–172. 62

Dorigo, M. and Gambardella, L. M. (1996). Ant colony system: A cooperative learning approach to the traveling salesman problem. Technical Report TR/IRIDIA/1996-5, Université Libre de Bruxelles. 3, 7, 8, 61, 62, 64, 102, 103

Dorigo, M. and Socha, K. (2006). An introduction to ant colony optimization. Technical Report TR/IRIDIA/2006-010, IRIDIA, Institut de Recherches Interdisciplinaires et de Développements en Intelligence Artificielle , Université Libre de Bruxelles. 61

Dorigo, M. and Stützle, T. (2009). Ant colony optimization: Overview and recent advances. Technical Report TR/IRIDIA/2009-013, IRIDIA, Institut de Recherches Interdisciplinaires et de Développements en Intelligence Artificielle, Université Libre de Bruxelles. 7, 61

Dorogovtsev, S. (2010). *Lectures on Complex Networks*. Oxford University Press. 24

Dorogovtsev, S. N., Goltsev, A. V., and Mendes, J. F. F. (2002). Pseudofractal scale-free web. *Phys. Rev. E*, 65:066122. 32

Dorogovtsev, S. N. and Mendes, J. F. F. (2003). *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press, USA. 24

Du, H., Feldman, M. W., Li, S., and Jin, X. (2007). An algorithm for detecting community structure of social networks based on prior knowledge and modularity. *Complexity*, 12(3):53–60. 56

Duch, J. and Arenas, A. (2005). Community detection in complex networks using extremal optimization. *Phys. Rev. E*, 72:027104. 57

Duckstein, L., Bartels, P., and Weber, J. (1988). Organization of a knowledge base by q-analysis. *Applied Mathematics and Computation*, 26(4):289–301. 46, 47

Duckstein, L. and Nobe, S. A. (1997). Q-analysis for modeling and decision making. *European Journal of Operational Research*, 103(3):411–425. 47

Dumais, S. and Chen, H. (2000). Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA. ACM. 19

Erdös, P. and Rényi, A. (1959). On random graphs I. *Publicationes Mathematicae Debrecen*, 6:290–297. 2, 27, 51

Erdös, P. and Rényi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5:17–61. 2, 27, 51

Erdös, P. and Rényi, A. (1961). On the strength of connectedness of a random graph. *Acta Mathematica Scientia Hungary*, 12(1):261–267. 27, 51

Euler, L. (1741). Solvtio problematis ad geometriam sitvs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, 8(53):128–140. 23, 51

Feng, A. and Allan, J. (2009). Incident threading for news passages. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 1307–1316, New York, NY, USA. ACM. 20

Fix, E. and Hodges, J. (1951). Discriminatory analysis, nonparametric discrimination: Consistency properties. Technical Report 4, USAF School of Aviation Medicine, Randolph Field, Texas. 18

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3-5):75–174. 52, 58, 60

Fortunato, S. and Barthelemy, M. (2006). Resolution limit in community detection. *physics/0607100*. Proc. Natl. Acad. Sci. USA 104 (1), 36-41 (2007). 54

Freeman, L. C. (1980). Q-analysis and the structure of friendship networks. *International Journal of Man-Machine Studies*, 12(4):367–378. 38

Fuhr, N., Hartmann, S., Lustig, G., Schwantner, M., Tzeras, K., and Knorz, G. (1991). Air/x - a rule-based multistage indexing system for large subject fields. In *Proceedings of RIAO'91*, pages 606–623. 18

Gambardella, L. M. and Dorigo, M. (1995). Ant-q: A reinforcement learning approach to the travelling salesman problem. In Kaufman, M., editor, *Proceedings of the ML-95, Twelfth Intern. Conf. on Machine Learning*, pages 252–260. 61

Ganguly, N., Deutsch, A., and Mukherjee, A. (2009). *Dynamics on and of Complex Networks: Applications to Biology, Computer Science, and the Social Sciences*. Springer. 35

Gatrell, A. C. (1981). On the structure of urban social areas: Explorations using q-analysis. *Transactions of the Institute of British Geographers*, 6(2):pp. 228–245. 45

Gfeller, D. and De Los Rios, P. (2008). Spectral coarse graining and synchronization in oscillator networks. *Phys. Rev. Lett.*, 100:174104. 60

Girvan, M. and Newman, M. (2002). Community structure in social and biological networks. *cond-mat/0112110*. Proc. Natl. Acad. Sci. USA 99, 7821-7826 (2002). 51, 75, 98

Gould, P. (1980). Q-analysis, or a language of structure: an introduction for social scientists, geographers and planners. *International Journal of Man-Machine Studies*, 13(2):169–199. 43

Gould, P. and Johnson, J. H. (1980). The content and structure of international television flows. *Communication*, 5:43–63. 38, 39, 46

Gross, T. (2009). Adaptive networks: State-topology coevolution. Online http://adaptive-networks.wikidot.com/. 7, 30

Gross, T. and Blasius, B. (2008). Adaptive coevolutionary networks: a review. *Journal of The Royal Society Interface*, 5(20):259 –271. 2, 7, 31, 32, 34

Gross, T. and Sayama, H., editors (2009). *Adaptive Networks: Theory, Models and Applications*. Springer. 7, 30, 35

Grüter, C., Schürch, R., Czaczkes, T. J., Taylor, K., Durance, T., Jones, S. M., and Ratnieks, F. L. W. (2012). Negative feedback enables fast and flexible collective decision-making in ants. *PLoS ONE*, 7(9):e44501. 64

Guimerà, R. and Nunes Amaral, L. A. (2005). Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900. 57

Guimerà, R., Sales-Pardo, M., and Amaral, L. A. N. (2004). Modularity from fluctuations in random graphs and complex networks. *Phys. Rev. E*, 70:025101. 57

Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena, CA USA. 147

Hamamoto, M., Kitagawa, H., Pan, J., and Faloutsos, C. (2005). A comparative study of feature Vector-Based topic detection schemes a comparative study of feature Vector-Based topic detection schemes. In *Web Information Retrieval and Integration, 2005. WIRI '05. Proceedings. International Workshop on Challenges in*, pages 122–127. IEEE. 18, 19

Hamill, L. and Gilbert, N. (2008). A simple but more realistic agent-based model of a social network. In *Conf. European Social Simulation Assoc. (ESSA'08)*. 2

Hammersley, J. M. (1957). Percolation processes. *Mathematical Proceedings of the Cambridge Philosophical Society*, 53(03):642–645. 32

Handl, J. and Meyer, B. (2002). Improved ant-based clustering and sorting in a document retrieval interface. In Guervós, J., Adamidis, P., Beyer, H.-G., Schwefel, H.-P., and Fernández-Villacañas, J.-L., editors, *Parallel Problem Solving from Nature - PPSN VII*, volume 2439 of *Lecture Notes in Computer Science*, pages 913–923. Springer Berlin / Heidelberg. 67

Hatzivassiloglou, V., Gravano, L., and Maganti, A. (2000). An investigation of linguistic features and clustering algorithms for topical document clustering. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 224–231, Athens, Greece. 17

Hopcroft, J., Khan, O., Kulis, B., and Selman, B. (2003). Natural communities in large linked networks. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '03, pages 541–546, New York, NY, USA. ACM. 59

Iredi, S., Merkle, D., and Middendorf, M. (2001). Bi-criterion optimization with multi colony ant algorithms. In Zitzler, E., Thiele, L., Deb, K., Coello Coello, C., and Corne, D., editors, *Evolutionary Multi-Criterion Optimization*, volume 1993 of *Lecture Notes in Computer Science*, pages 359–372. Springer Berlin Heidelberg. 65

Ishida, Y., Adachi, N., and Tokumaru, H. (1985). A topological approach to failure diagnosis of large-scale systems. *IEEE Trans. SMC*, 15(3):327–333. 46

Jacobson, T. (2003). Visualizing Information Seeking With Q-analysis Q-analysis. *Paper presented at a non-divisional workshop held at the meeting of the International Communication Association, San Diego, CA.* 41

Jacobson, T. L. and Yan, W. (1998). Q-analysis techniques for studying communication content. *Quality & Quantity*, 32:93–108. 10.1023/A:1004255229611. 48, 49

Jafar, O. M. and Sivakumar, R. (2010). Ant-based clustering algorithms: A brief survey. *International Journal of Computer Theory and Engineering*, 2(5). 7, 66

Jo, T., Seo, J., and Kim, H. (2000). *Topic Spotting on News Articles with Topic Repository by Controlled Indexing*. Intelligent Data Engineering and Automated Learning ,Äî IDEAL 2000. Data Mining, Financial Engineering, and Intelligent Agents. Springer Berlin Heidelberg, Berlin, Heidelberg. 1983. 17

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *Machine Learning ECML98*, 1398(23):137–142. 18

Johnson, J. H. (1976). The q-analysis of road intersections. *International Journal of Man-Machine Studies*, 8(5):531–548. 7, 76

Johnson, J. H. (1978a). A Q-analysis of television programmes. *International Journal of ManMachine Studies*, 10(4):461–479. 46

Johnson, J. H. (1978b). A q-analysis of television programmes. *International Journal of Man-Machine Studies*, 10(4):461 – 479. 7

Johnson, J. H. (1981a). The q-analysis of road traffic systems. *Environment and Planning B: Planning and Design*, 8(2):141–189. 45, 46

Johnson, J. H. (1981b). Some structures and notation of Q-analysis. *Environment And Planning B*, 8:73–86. 39, 96

Johnson, J. H. (1982). q-transmission in simplicial complexes. *International Journal of Man-Machine Studies*, 16(4):351 – 377. 45

Johnson, J. H. (1983). A survey of q-analysis, part 1: The past and present. In *Proceedings of the Seminar on Q-analysis and the Social Sciences, Universty of Leeds*. xiii, 3, 39, 40, 41, 44

Johnson, J. H. (1990). Expert q-analysis. *Environment and Planning B: Planning and Design*, 17(2):221–244. 49

Johnson, J. H. (2005). Multidimensional multilevel networks in the science of the design of complex systems. In Johnson, J., editor, *ECCS 2005 Satellite Workshop: Embracing Complexity in Design*, volume ECCS 2005 Satellite Workshop: Embracing Complexity in Design. 96

Johnson, J. H. (2006a). Can complexity help us better understand risk? *Risk Managment*, 8(4):227–267. 96

Johnson, J. H. (2006b). Hypernetworks for reconstructing the dynamics of multilevel systems. In *European Conference on Complex Systems 2006*. 2

Kim, D. and Oh, A. (2011). Topic chains for understanding a news corpus. In *Proceedings of the 12th international conference on Computational linguistics and intelligent text processing - Volume Part II*, CICLing'11, pages 163–176, Berlin, Heidelberg. Springer-Verlag. 21

Kohut, A. and Remez, M. (2008). Internet overtakes newspapers as news outlet. http://www.people-press.org/2008/12/23/internet-overtakes-newspapers-as-news-outlet/. xiii, 4, 6

Kuramoto, Y. (1975). Self-entrainment of a population of coupled nonlinear oscillators. *International symposium on mathematical problems in theoretical physics, Lecture notes in Physics*, 39:420–422. 60

Kuramoto, Y. (1984). *Chemical oscillations, waves, and turbulence*. Springer-Verlag, New York. 60

Lavrenko, V., Allan, J., DeGuzman, E., LaFlamme, D., Pollard, V., and Thomas, S. (2002). Relevance models for topic detection and tracking. In *Proceedings of the second international conference on Human Language Technology Research*, HLT '02, pages 115–121, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. 20

Lee, M., Wang, W., and Yu, H. (2006). Exploring supervised and unsupervised methods to detect topics in biomedical text. *BMC bioinformatics*, 7:140–140. 17

Lewis, D. D. and Ringuette, M. (1994). A comparison of two learning algorithms for text categorization. In *In Third Annual Symposium on Document Analysis and Information Retrieval*, pages 81–93. 18

Lin, S. (1965). Computer solutions of the traveling salesman problem. *Bell System Technical Journal*, 44(10):2245–2269. 62

Lin, S.-h. and Ho, J.-m. (2002). Discovering informative content blocks from web documents. *Knowledge Creation Diffusion Utilization*, pages 1–9. 79

Liu, S.-W. and Chang, H.-T. (2013). A topic detection and tracking system with tf-density. In Gaol, F. L., editor, *Recent Progress in Data Engineering and Internet Technology*, volume 156 of *Lecture Notes in Electrical Engineering*, pages 115–120. Springer Berlin Heidelberg. 21

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proc. 5th Berkeley Symp. Math. Stat. Probab., Univ. Calif. 1965/66, 1, 281-297 (1967). 19, 83

McKeown, K. R., Barzilay, R., Evans, D., Hatzivassiloglou, V., Klavans, J. L., Nenkova, A., Sable, C., Schiffman, B., and Sigelman, S. (2002). Tracking and summarizing news on a daily basis with columbia's newsblaster. In *Proceedings of the second international conference on Human Language Technology Research*, HLT '02, pages 280–285, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. 20

Meilă, M. (2007). Comparing clusterings—an information based distance. *J. Multivar. Anal.*, 98(5):873–895. 7, 75, 77, 78, 79

Melville, B. (1976). Notes on the civil applications of mathematics. *International Journal of Man-Machine Studies*, 8(5):501–515. 43

Miao, Y. and Qiu, X. (2009). Hierarchical centroid-based classifier for large scale text classification. In *Large Scale Hierarchical Text classification (LSHTC) Pascal Challenge*. 18

Milgram, S. (1967). The small-world problem. *Psychology Today*, 1(1):60–67. 28

Montgomery, J. and Randall, M. (2002). Anti-pheromone as a tool for better exploration of search space. In *Proceedings of the Third International Workshop on Ant Algorithms*, ANTS '02, pages 100–110, London, UK, UK. Springer-Verlag. 64

Newman, M. (2003). The structure and function of complex networks. *SIAM Review*, 45(2):167–256. 32

Newman, M. (2004). Fast algorithm for detecting community structure in networks. *Phys. Rev. E*, 69:066133. 56

Newman, M. (2006a). Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E*, 74:036104. 58

Newman, M. (2006b). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582. 53, 58, 75

Newman, M., Barabasi, A.-L., and Watts, D. J. (2006). *The Structure and Dynamics of Networks:*. Princeton University Press, 1 edition. 27

Newman, M. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113. 53, 55

Ng, H. T., Goh, W. B., and Low, K. L. (1997). Feature selection, perceptron learning, and a usability case study for text categorization. *SIGIR Forum*, 31(SI):67–73. 18

Nigam, K., Lafferty, J., and McCallum, A. (1999). Using maximum entropy for text classification. In *IJCAI-99 Workshop on Machine Learning for Information Filtering*, pages 61–67. 17

Palla, G., Barabasi, A.-L., and Vicsek, T. (2007). Quantifying social group evolution. *Nature*, 446(7136):664–667. 59

Palla, G., Derényi, I., Farkas, I., and Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435:814–8. PMID: 15944704. 51, 59

Pantel, P. and Lin, D. (2001). A statistical corpus-based term extractor. In *Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence*, pages 1–10. Springer-Verlag. 17

Pastor-Satorras, R. and Vespignani, A. (2001). Epidemic spreading in scale-free networks. *Phys. Rev. Lett.*, 86:3200–3203. 32

Pikovsky, A., Rosenblum, M., and Kurths, J. (2003 [2001]). *Synchronization: A Universal Concept in Nonlinear Sciences*. Cambridge Nonlinear Science Series. Cambridge University Press, 3rd edition. 59

Pool, I. d. S. and Kochen, M. (1978 [1958]). Contacts and influence. *Social Networks*, 1(1):5–51. 28

Porter, M. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137. 145

Pujol, J. M., Béjar, J., and Delgado, J. (2006). Clustering algorithm for determining community structure in large networks. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 74(1 Pt 2):016107. PMID: 16907151. 56

Qi, X. and Davison, B. D. (2009). Web page classification: Features and algorithms. *ACM Comput. Surv.*, 41(2):12:1–12:31. 145

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. 90

Ramos, V., Fernandes, C., and Rosa, A. C. (2006). On self-regulated swarms, societal memory, speed and dynamics. In *Artificial Life X: Proceedings of the Tenth International Conference on the Simulation and Synthesis of Living Systems*, pages 393–399. MIT Press. 113

Ramos, V. and Merelo, J. J. (2002). Self-organized stigmergic document maps: Environment as a mechanism for context learning. In *Proceddings of the AEB*, Mérida, Spain. 66

Ramos, V., Rodrigues, D. M. S., and Louçã, J. (2011). Spatio-temporal dynamics on co-evolved stigmergy. In Thurner, S., editor, *8th European Conference on Complex Systems*, Vienna, Austria. 8, 101

Ramos, V., Rodrigues, D. M. S., and Louçã, J. (2013). Second order swarm intelligence. In Pan, J.-S., Polycarpou, M., Woźniak, M., Carvalho, A., Quintián, H., and Corchado, E., editors, *HAIS'13. 8th International Conference on Hybrid Artificial Intelligence Systems*, volume 8073 of *Lecture Notes in Computer Science*, pages 411–420. Springer Berlin Heidelberg, Salamanca, Spain. 7, 8, 73, 101

Robinson, E. J., Jackson, D., Holcombe, M., and Ratnieks, F. L. (2007). No entry signal in ant foraging (hymenoptera: Formicidae): new insights from an agent-based model. *Myrmecological News*, 10(120). 64

Robinson, E. J. H., Jackson, D. E., Holcombe, M., and Ratnieks, F. L. W. (2005). Insect communication: 'no entry' signal in ant foraging. *Nature*, 438(7067):442–442. 64

Rodrigues, D. M. S. (2010). The observatorium – the structure of news: topic monitoring in online media with mutual information. In Louçã, J., editor, *Proceedings of the European Conference on Complex Systems*, Lisbon. Complex Systems Society. 7, 9, 73, 75, 77

Rodrigues, D. M. S. (2013). Identifying news clusters using q-analysis and modularity. In Diaz-Guilera, A., Arenas, A., and Corral, Á., editors, *Proceedings of the European Conference on Complex Systems 2013*, Barcelona. 8, 9, 73, 89

Rodrigues, D. M. S., Louçã, J., and Ramos, V. (2011). From standard to second-order swarm intelligence phase-space maps. In Thurner, S., editor, *8th European Conference on Complex Systems*, Vienna, Austria. 7, 8, 101

Rodrigues, D. M. S. and Louçã, J. (2009). Mutual information to assess structural properties in dynamic networks. In *ECCS09 conference proceedings. Warwick 2009*. 7, 9, 73, 75, 77

Rohlf, T. and Bornholdt, S. (2009). Self-Organized criticality and adaptation in discrete dynamical networks. In Gross, T. and Sayama, H., editors, *Adaptive Networks*, volume 51 of *Understanding Complex Systems*, pages 73–106. Springer Berlin / Heidelberg. 10.1007/978-3-642-01284-6_5. 36

Ruiz, M. E. and Srinivasan, P. (2002). Hierarchical text categorization using neural networks. *Inf. Retr.*, 5(1):87–118. 18

Saavedra, S., Efstathiou, J., and Reed-Tsochas, F. (2007). Identifying the underlying structure and dynamic interactions in a voting network. *Physica A*, (377):672–688. 32

Schnettler, S. (2009). A structured overview of 50 years of small-world research. *Social Networks*, 31(3):165–178. 28, 29

Schoonderwoerd, R., Bruten, J. L., Holland, O. E., and Rothkrantz, L. J. M. (1996). Ant-based load balancing in telecommunications networks. *Adapt. Behav.*, 5(2):169–207. 64

Schuetz, P. and Caflisch, A. (2008a). Efficient modularity optimization by multistep greedy algorithm and vertex mover refinement. *Phys. Rev. E*, 77:046112. 56

Schuetz, P. and Caflisch, A. (2008b). Multistep greedy algorithm identifies community structure in real-world and computer-generated networks. *Phys. Rev. E*, 78:026112. 56

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47. 18

Solé, R. V., Corominas-Murtra, B., Valverde, S., and Steels, L. (2010). Language networks: Their structure, function, and evolution. *Complexity*, 15(6):20–26. 18

Solomonoff, R. (1952). An exact method for the computation of the connectivity of random nets. *The bulletin of mathematical biophysics*, 14:153–157. 27

Solomonoff, R. and Rapoport, A. (1951). Connectivity of random nets. *Bulletin of Mathematical Biology*, 13:107–117. 2, 27

Steinhaus, H. (1957). Sur la division des corps matériels en parties. *Bull. Acad. Pol. Sci., Cl. III*, 4:801–804. 20

Travers, J. and Milgram, S. (1969). An experimental study of the small world problem. *Sociometry*, 32(4):425–443. 2, 28

Tzeras, K. and Hartmann, S. (1996). Automatic indexing based on bayesian inference networks. In Korfhage, R., Rasmussen, E. M., and Willett, P., editors, *16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 22–34, Pittsburgh, PA USA. ACM. 18

van Rijsbergen, C., Robertson, S., and Porter, M. (1980). New models in probabilistic information retrieval. Technical Report 5587, British Library, London, England. 145

Vapnik, V. (1965 [1982]). *Estimation of Dependences Based on Empirical Data: Springer Series in Statistics (Springer Series in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA. 18

Vizine, A. L., de Castro, L. N., and Gudwin, R. R. (2005). Text document classification using swarm intelligence. In *Proceedings of the International Conference on Integration of Knowledge Intensive Multi-Agent Systems*. 67

Wakita, K. and Tsurumi, T. (2007). Finding community structure in mega-scale social networks: [extended abstract]. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 1275–1276, New York, NY, USA. ACM. 56

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Number 8 in Structural Analysis in the Social Sciences. Cambridge University Press, 1 edition. 44

Watts, D. and Strogatz, S. (1998). Collective dynamics of small-world networks. *Nature*, 293:420–442. 2, 28, 32

Weninger, T. and Hsu, W. (2008). Text extraction from the web via text-to-tag ratio. In *Database and Expert Systems Application, 2008. DEXA'08. 19th International Workshop on*, pages 23–28. IEEE. 17, 79, 146

Wiener, E., Pedersen, J., and Weigend, A. (1995). A neural network approach to topic spotting. In *Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval*. 18

Xiang, B., Chen, E.-H., and Zhou, T. (2009). Finding community structure based on subgraph similarity. In Fortunato, S., Mangioni, G., Menezes, R., and Nicosia, V., editors, *Complex Networks*, volume 207 of *Studies in Computational Intelligence*, pages 73–81. Springer Berlin / Heidelberg. 56

Yang, T., Jin, R., Chi, Y., and Zhu, S. (2009). Combining link and content for community detection: a discriminative approach. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 927–936, New York, NY, USA. ACM. 21

Yang, Y. and Chute, C. G. (1994). An example-based mapping method for text categorization and retrieval. *ACM Trans. Inf. Syst.*, 12(3):252–277. 18

Yang, Y. and Liu, X. (1999). A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 42–49. ACM. 17, 18

Ye, Z., Hu, S., and Yu, J. (2008). Adaptive clustering algorithm for community detection in complex networks. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 78(4 Pt 2):046115. PMID: 18999501. 56

Zhang, Z.-k., Zhou, T., and Zhang, Y.-c. (2010). Personalized recommendation via integrated diffusion on user-item- tag tripartite graphs. *Physica A: Statistical Mechanics and its Applications*, 389(1):179–186. 17

Zhou, Y., Cheng, H., and Yu, J. X. (2009). Graph clustering based on structural/attribute similarities. *Proc. VLDB Endow.*, 2(1):718–729. 21

Part III

Appendices

Appendix A

Theseus - A crawling and analysis platform of online content

The high volatility of online media news make it very difficult to track news as they are constantly produced around the world. For this it was necessary to use some structure to gather and fetch news from the internet and store them. The purpose of this project did not need a full fledged web crawler, so this development lead to a python package to crawl and process the webpages of online media, taking advantage of the typical structure of these news sources. This then lead to the development of Theseus[22], a set of python programs and modules for the retrieval, storage, cleaning and classification/clustering of news from online media.

A.1 Rationality behind Theseus

The objective of studying the structure of texts published online demands that different approaches and techniques be used according to the different channels used by the publishing sources (newspapers, blogs, social networks). This in turn requires that an application that aims at doing an integrated and global analysis has to be flexible and continuously adapted to the permanent publication format changes, API[23] changes, and services published by each of the contents channels. Due to this volatile aspect, it was decided to program the tools necessary so they could be tailored to the needs of this work, instead of opting for *off-the-shelf* text analysis tools. That option would constrain the research to what those tools were capable of and would also require the learning of a multitude of tools to adapt the workflow of each one to the analysis methodology developed. This has not eliminated the use of *third-party* tools completely and they were used where one wanted to use a particular aspect of it that was clearly documented, easy to integrate with this tool, and would not interfere with the methods and structure developed. The dependencies of Theseus are described in subsection A.3.3.

It was of interest to develop tools released under an open source licences, so they could

[22] Theseus is available as an open source project at https://github.com/sixhat/theseus and documentation can be found at http://theobservatorium.com/html/

[23] Application Programming Interface

be validated by other users; that they were highly modular for easy of reuse in other tools; and that they allowed for a quick development cycle. It was required that the tool allowed the chaining of different modules to be developed or integrated from other software, as to allow the development of complex analysis chains. For these reasons of these when it was not possible to find an external software module that fitted these ideas, one opted to develop from scratch the tool needed to do the tasks, avoiding the use of external dependencies.

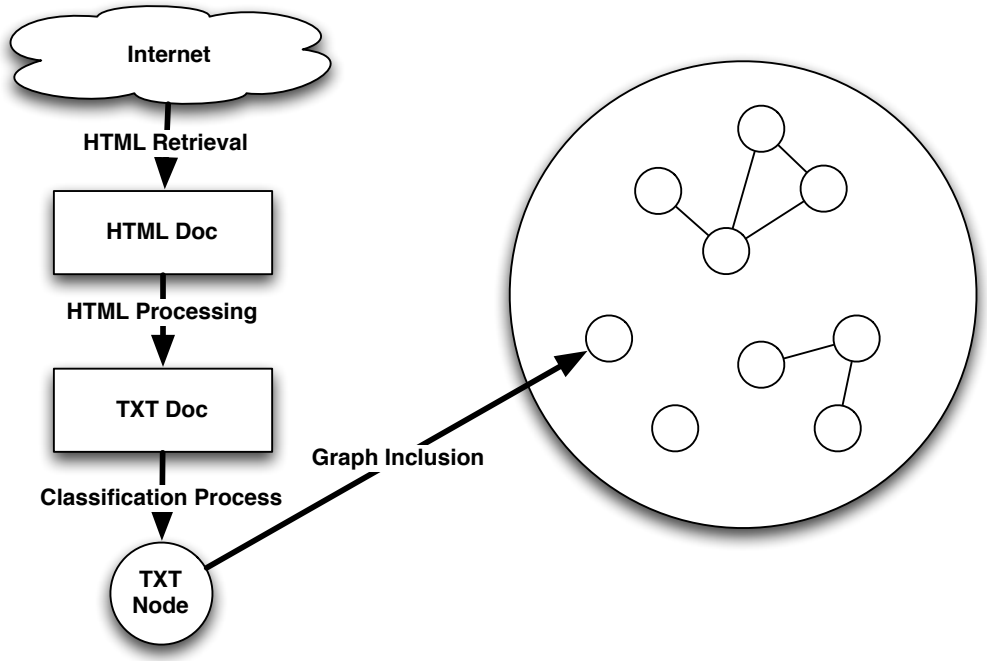


Figure A.1: Example of a processing sequence used in Theseus

An example of a typical workflow implemented can be seen in Figure A.1. In this example, a document is downloaded from the internet, and pre-processed for text extraction, it is then put in a structure representing a network node and lastly it is included in a graph according the chosen algorithm. I wanted these steps to act as independent units in a modular way so that the system would allow the quick replacement of on these operations with others. Because of this, the programming language **Python**[24] was used, mainly because of its dynamical aspect, allowing the tools to be quickly developed and leaving to the user the final assembly of an analysis chain according to his needs.

Python was not the only programming language used in the development of Theseus. **R**, **C** or **Java** were used for specific tasks that leveraged their particular strengths. In any case the core of Theseus is developed in Python and it acts as the *glue* that connects the different components together.

[24]Website: http://www.python.org/

A.2 Document preprocessing and representation

Qi and Davison (2009) assert that text classification is a general problem that is different from web content classification:

> First, traditional text classification is typically performed on "structured corpora with well-controlled authoring styles" (...), while web collections do not have such a property. Second, web pages are semi-structured documents in HTML, so that they may be rendered visually for users. Although other document collections may have embedded information for rendering and/or a semi-structured format, such markup is typically stripped for classification purposes. Finally, web documents exist within a hypertext, with connections to and from other documents. While not unique to the web (consider for example the network of scholarly citations), this feature is central to the definition of the web, and is not present in typical text classification problems.

To be able to process the collected documents, it is necessary that they are pre-processed. This pre-processing phase is usually made of multiple sequential steps, though not all of those steps are implemented for every corpus, as a way of preparing the documents for the machine learning algorithm chosen.

In this pre-processing phase, the HTML documents must be cleaned of all HTML markup, Javascript, and embedded objects like flash. The document is then in a text format has to be treated so as to deal with numbers, punctuation, and word capitalisation. It is then necessary to remove stopwords. Stopwords are those words that are very common in a specific language and that do not convey important information. This helps reduced the dimensionality of the document without significant loss in accuracy. After this phase, usually a stemming of the terms found in the document is performed as way to reduce the document to the stemmed version. For this a Porter algorithm is used for English language documents while others are available for other languages (Porter, 1980; van Rijsbergen et al., 1980). After constructing this stemmed document, usually there is a process of term weighting. Several weighting methods are available with term frequency (TF) and term frequency inverse document frequency (TFIDF) being two of the most common. After this process in some cases it is necessary to reduce the dimensionality of the problem by reducing the number of terms analysed. These pre-processing tasks can be accomplished easily by the processor module of Theseus.

A.3 The components of Theseus

Theseus is not a monolithic traditional program but a collection of libraries, scripts, and small programs that can be used together to build a system tailored to the needs of the problem in hand. Theseus mixes several technologies and languages, mainly python, javascript and bash. It is mainly aimed at unix type systems as parts of it are targeted to deployment on linux systems, but a port to other systems is possible.

A.3.1 The crawler

This Python module is responsible for parsing stories from RSS feeds and download the corresponding news items to text files that can then later be processed in other ways.

A journal is scheduled to be download via a linux cronjob that runs at a predetermined time interval and are responsible for:

1. Fetchs and parses the XML RSS file from the journal for individual news items.
2. For each news item in the RSS feed not already downloaded, download the original source of the news item.
3. Store the downloaded files in a flat file storage system for further processing.

A.3.2 The processor

This Python module is responsible for processing text, calculating features, summarising text, etc... A brief summary of the features present in the processor module of Theseus is given:

The Theseus processor defines a set of high level objects that make the text manipulation task easier. As it aims to work with document files the basic object is **DocNode** and is the basic structure that olds each document in a corpus.

All **DocNode**s present in a newspaper can then be aggregated in an object of class **Channel**.

Each **DocNode** is then composed of a collection of objects of the class **Sentence** where low level manipulation takes place.

These objects (**Channel, DocNode, Sentece**) all possess several methods that allow commons tasks to be performed. These include the cleaning of text of unauthorised characters, the creation of word lists, the removal of stopwords, etc..

Also, the processor module includes a set of methods to work on this collection of documents, like statistical methods to calculate histograms of words on a **Channel**, calculation of document frequency of words in several variants: inverse document frequency (idf), document term frequency (dtf), log term frequency (logtf), log term frequency-inverse document frquency (logtf.idf), normalized frequency (normF), term frquency (tf), term frequency-inverse document frequency (tf.idf) and term frequency-proportional document frequency (tf*pdf).

The processor module also includes HTML to Text converters and has implemented the Text 2 Tag ratio proposed by Weninger and Hsu (2008).

A.3.3 Integration with existing tools and dependencies

Although designed from scratch to avoid dependencies, the growth in Theseus functionality made it necessary to introduce some dependencies on external libraries, namely:

- NetworkX - This is a Python language software package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks available from http://networkx.lanl.gov (Hagberg et al., 2008)
- Universal feedparser - RSS feed parser for parsing the XML files retrieved from the web journals from http://packages.python.org/feedparser/
- Matplotlib - Plotting and visualisation libraries from http://matplotlib.sourceforge.net
- Scipy and Numpy - Scientific numerical libraries from http://www.scipy.org
- D3.js - This library is used for the visualisation of graph structure in webpages available from http://mbostock.github.com/d3.

Besides these libraries and tools, Theseus was designed to run in a linux environment and therefore is integrated with the underlying system through bash scripts, cronjobs, etc. . .

Appendix B

News topics identified by the *Q*-analysis filtering method

Sample of the List of topics identified by the algorithm of Chapter 10. A full listing of the topics would be extensive so a copy of this full appendix is available in digital format online at http://www.davidrodrigues.org/pdfs/phd/appendix-B.pdf and in the accompanying CD to this thesis.

Topic

- Academic linked to Gaddafi's fugitive son leaves LSE - Education - The Guardian
- Gaddafi donation to LSE may have come from bribes, inquiry finds - Education - The Guardian

Topic

- Berlusconi denies quit claims and stakes future on economic reform vote - World news - The Guardian
- G20 summit fails to allay world recession fears - World news - guardian.co.uk
- Cameron warned his eurozone stance risks forcing two-speed Europe - Business - The Guardian
- Greece throws euro bailout into fresh crisis - Business - The Guardian
- European debt crisis is worst time since second world war, says Angela Merkel - Business - guardian.co.uk
- G20: Europe faces the nightmare of a euro breakup - World news - guardian.co.uk
- Markets hit by European debt crisis as threat of eurozone breakup persists - Business - guardian.co.uk
- Greece prepares for departure of George Papandreou - World news - guardian.co.uk
- Spain's debt crisis worsens as country begins month of post-election limbo - World news - guardian.co.uk
- Silvio Berlusconi bows out after Italian MPs vote for savage cuts - World news - The Observer
- Eurozone crisis: Cameron and Merkel agree framework for EU negotiations - Business - The Guardian

- Greek government teeters on brink of collapse in wake of referendum plan - Business - The Guardian
- G20 summit: Cameron will urge EU to flesh out bailout deal - Business - guardian.co.uk
- The euro will survive – and Britain will join, says Michael Heseltine - World news - guardian.co.uk
- Chaos in Greece amid battle to form a 'government of national salvation' - World news - The Observer
- Papandreou out as Greek leaders agree unity government deal - World news - The Guardian
- Euro stability more important than Greece, says Angela Merkel - Business - The Guardian
- Eurozone crisis stokes tension between Britain and Germany - Business - The Guardian
- Central banks step in to stave off new credit crunch - Business - The Guardian
- IMF denies reports of Italy bailout deal - Business - guardian.co.uk
- Eurozone debt crisis: EU members line up to demand ECB intervention - Business - The Guardian
- Eurozone looks to IMF as contagion spreads - Business - The Guardian
- Greek PM wins backing for referendum - European debt crisis live - Business - guardian.co.uk
- 'Tobin tax' would hit City of London with missile, says John Major - Business - The Guardian
- Greece may leave euro, leaders admit - Business - The Guardian
- Lucas Papademos to lead Greece's interim coalition government - World news - guardian.co.uk
- Global recession grows closer as G20 summit fails - Business - The Guardian
- Eurozone crisis gives Britain a chance to redraw, says David Cameron - Business - The Guardian
- Greek government on brink of collapse over debt crisis - World news - guardian.co.uk
- Global recession grows closer as G20 summit fails - Business - The Guardian
- Eurozone finance ministers turn to IMF to help bailout fund - Business - guardian.co.uk
- G20 leaders discuss eurozone crisis: live coverage - Politics - guardian.co.uk
- Greek PM struggles to find way out of crisis - World news - guardian.co.uk
- Better-than-expected UK growth fails to lift economists' outlook - Business - The Guardian
- Greek leaders agree to unity government as future hangs in balance - World news - guardian.co.uk
- Silvio Berlusconi to bow out after Italian MPs vote for savage cuts - World news - The Observer

- Euro debt crisis: Greek PM George Papandreou to resign - World news - guardian.co.uk
- UK on the brink of double-dip recession, warns OECD - Business - guardian.co.uk

Topic

- Egypt protests: elections in doubt - live updates - World news - guardian.co.uk
- Libya, Syria and Middle East unrest - live updates - World news - guardian.co.uk
- Egypt and Syria protests - live updates - World news - guardian.co.uk
- Libyan PM snubs Islamists with cabinet to please western backers - World news - The Guardian
- Egypt: The return to Tahrir Square- live updates - World news - guardian.co.uk
- Egypt: protests continue in runup to elections – live - World news - guardian.co.uk
- Tahrir Square: police clash with protesters live coverage - World news - guardian.co.uk
- Syria crisis: Assad given three days to end bloodshed - live updates - World news - guardian.co.uk
- Syria: Arab League meets as defectors attack - live updates - World news - guardian.co.uk
- Egypt protests: 'Friday of the last chance' - live updates - World news - guardian.co.uk
- Saif al-Islam Gaddafi captured in Libyan desert as he tries to flee country - World news - The Observer
- Egypt elections - live updates - World news - guardian.co.uk
- Arab League's Syria plan put to the test - live updates - World news - guardian.co.uk
- Egypt election results and Turkey slaps sanctions on Syria - live updates - World news - guardian.co.uk
- Arab League to reveal Syria peace plan - live updates - World news - guardian.co.uk
- Arab League to reveal Syria peace plan - World news - guardian.co.uk

Topic

- China space station modules dock in historic rendezvous - World news - guardian.co.uk
- Chinese spacecraft lands safely in Mongolia - World news - guardian.co.uk
- China hits back over US claims of satellite hacking - Technology - guardian.co.uk
- China sends unmanned craft into space - World news - guardian.co.uk
- William Hague backs off criticising China over cyber attacks - Technology - The Guardian

Topic

- Ed Miliband: politicians must listen to the St Paul's Cathedral protesters - UK news - The Observer
- Vince Cable: I sympathise with Occupy London protesters at St Paul's - UK news - guardian.co.uk
- Occupy London faces eviction: live Q&A on how to proceed - Hannah Borno - Comment is free - guardian.co.uk
- St Paul's suspends legal action against protesters – live coverage - UK news - guardian.co.uk

- Occupy London: eviction bid cites desecration, defecation and drugs - UK news - The Guardian
- Police arrest EDL members to 'avert planned attack' in London - UK news - The Guardian
- St Paul's Cathedral dean resigns over Occupy London protest row - UK news - The Guardian
- Veterans join Occupy protest as St Paul's canon shows support - UK news - The Observer
- Occupy London protesters take over empty UBS bank offices - UK news - The Guardian
- Occupy London: live coverage of St Paul's protests - UK news - guardian.co.uk
- Occupy protesters accuse Boris Johnson of defending the rich - Politics - guardian.co.uk
- Bishop of London moves to dissociate church from St Paul's legal action - UK news - guardian.co.uk
- St Paul's seeks new direction and suspends legal action - UK news - The Guardian
- Occupy London faces renewed eviction attempt - UK news - guardian.co.uk
- Occupy London: archbishop of Canterbury backs new tax on banking - UK news - guardian.co.uk
- Occupy London protesters allowed to stay at St Paul's until the new year - UK news - The Guardian
- St Paul's study reveals financial workers think City traders are overpaid - Business - guardian.co.uk
- St Paul's and Corporation of London halt legal action against Occupy camp - UK news - guardian.co.uk
- Occupy London: St Paul's camp eviction hearing opens on 19 December - UK news - guardian.co.uk
- Occupy London protesters say they were asked to leave St Paul's services - UK news - The Guardian
- Police arrest EDL members to 'avert planned attack' in London - UK news - The Guardian
- Occupy London protesters 'will not obstruct remembrance events' - UK news - The Guardian
- Occupy London: eviction notices attached to tents outside St Paul's - UK news - guardian.co.uk

Appendix C

Hamiltonian paths with double pheromone ant colony system optimisation

Sample of the Hamiltonian path of topics identified by the algorithm of chapter 11. for the month of November 2011 of The Guardian. A full listing of the topics would be extensive but a copy of the this full appendix is available in digital format online at http://www.davidrodrigues.org/pdfs/phd/appendix-C.pdf and in the accompanying CD to this thesis.

As news follows a Hamiltonian path the first news in the following list is connected to the last news of the list to complete the path.

- Eurozone debt crisis live: Italian senate passes austerity law | Business | guardian.co.uk
- Greek leaders agree to unity government as future hangs in balance | World news | guardian.co.uk
- Lucas Papademos to lead Greece's interim coalition government | World news | guardian.co.uk
- The euro will survive – and Britain will join, says Michael Heseltine | World news | guardian.co.uk
- Eurozone bailout fund falls short of €1 trillion target | Business | The Guardian
- Euro debt crisis: Greek PM George Papandreou to resign | World news | guardian.co.uk
- Chaos in Greece amid battle to form a 'government of national salvation' | World news | The Observer
- Eurozone debt crisis: EU members line up to demand ECB intervention | Business | The Guardian
- Italy passes austerity measures – clearing way for Berlusconi to quit | Business | guardian.co.uk
- European debt crisis live: pressure mounts as finance ministers meet | Business | guardian.co.uk
- Greek PM Papandreou faces knife-edge confidence vote: live | Business | guardian.co.uk
- Greece may leave euro, leaders admit | Business | The Guardian
- World central banks launch co-ordinated action to fight financial crisis - live | Business | guardian.co.uk

- Global recession grows closer as G20 summit fails | Business | The Guardian
- Eurozone finance ministers turn to IMF to help bailout fund | Business | guardian.co.uk
- Global recession grows closer as G20 summit fails | Business | The Guardian
- Silvio Berlusconi bows out after Italian MPs vote for savage cuts | World news | The Observer
- Berlusconi denies quit claims and stakes future on economic reform vote | World news | The Guardian
- Mario Monti heralds new era for Italy as Berlusconi bows out | World news | The Guardian
- Eurozone looks to IMF as contagion spreads | Business | The Guardian
- European debt crisis is worst time since second world war, says Angela Merkel | Business | guardian.co.uk
- Cameron warned his eurozone stance risks forcing two-speed Europe | Business | The Guardian
- Eurozone crisis: Cameron and Merkel agree framework for EU negotiations | Business | The Guardian
- Silvio Berlusconi hints at comeback as Italy tries to form new government | World news | guardian.co.uk
- Silvio Berlusconi to bow out after Italian MPs vote for savage cuts | World news | The Observer
- Crucial vote for eurozone due in Italian senate | Business | The Guardian
- Silvio Berlusconi to resign after austerity vote | World news | guardian.co.uk
- Italy's borrowing costs keep on rising despite Berlusconi's promise to quit | Business | guardian.co.uk
- Silvio Berlusconi vows to resign as Italy's prime minister | World news | The Guardian
- European debt crisis live: Greece locked in coalition talks | Business | guardian.co.uk
- Eurozone crisis: Spain's election leaves markets on edge | Business | guardian.co.uk
- Eurozone debt crisis: Berlusconi faces crunch vote - live | Business | guardian.co.uk
- Eurozone crisis will hit UK hard, warns Cameron | Business | The Guardian
- Papandreou's coalition offer snubbed by opposition leader | World news | guardian.co.uk
- Papandreou out as Greek leaders agree unity government deal | World news | The Guardian
- Greek PM struggles to find way out of crisis | World news | guardian.co.uk
- G20 summit fails to allay world recession fears | World news | guardian.co.uk
- Spanish boom town that went bust | World news | The Observer
- Spain's debt crisis worsens as country begins month of post-election limbo | World news | guardian.co.uk

- France, Germany and Italy squash market hopes of ECB intervention | Business | guardian.co.uk
- Eurozone debt crisis: Cameron and Merkel put on united front | Business | The Guardian
- Eurobond plan sets Barroso on collision course with Merkel | Business | The Guardian
- G20: Europe faces the nightmare of a euro breakup | World news | guardian.co.uk
- Euro stability more important than Greece, says Angela Merkel | Business | The Guardian
- Greece prepares for departure of George Papandreou | World news | guardian.co.uk
- G20 leaders discuss eurozone crisis: live coverage | Politics | guardian.co.uk
- Eurozone crisis stokes tension between Britain and Germany | Business | The Guardian
- Eurozone crisis gives Britain a chance to redraw, says David Cameron | Business | The Guardian
- Angela Merkel: Europe's saviour – or biggest problem? | World news | The Guardian
- Spanish election: exit polls suggest landslide victory for People's party | World news | The Guardian
- Irish property tycoon Sean Quinn declares bankruptcy in UK | World news | guardian.co.uk
- 'Tobin tax' would hit City of London with missile, says John Major | Business | The Guardian
- G20 summit: Cameron will urge EU to flesh out bailout deal | Business | guardian.co.uk
- G20 summit: live coverage | World news | guardian.co.uk
- Greek PM wins backing for referendum - European debt crisis live | Business | guardian.co.uk
- Eurozone crisis: Sarkozy, Merkel and Monti meet after German debt auction - live | Business | guardian.co.uk
- European debt crisis live: Greek referendum throws markets into turmoil | Business | guardian.co.uk
- Greek crisis: finance minister breaks ranks over referendum - live | Business | guardian.co.uk
- Greek government teeters on brink of collapse in wake of referendum plan | Business | The Guardian
- Greece throws euro bailout into fresh crisis | Business | The Guardian
- Greek government on brink of collapse over debt crisis | World news | guardian.co.uk
- Afghan finance minister admits doubts over Kabul Bank's missing $1bn | World news | The Guardian
- US removes Afghanistan commander Peter Fuller for criticising Karzai | World news | guardian.co.uk

- US 'kill team' soldier Calvin Gibbs convicted of murdering Afghan civilians | World news | guardian.co.uk
- US 'kill team' trial: jury considers Calvin Gibbs verdict | World news | guardian.co.uk
- Obama Australia visit begins | World news | guardian.co.uk
- Obama tells Asia US 'here to stay' as a Pacific power | World news | guardian.co.uk
- China uneasy over US troop deal in Australia | World news | The Guardian
- China baby-trafficking ring is shut down | World news | guardian.co.uk
- Hillary Clinton: after Afghanistan and Iraq, Asia-Pacific is next | World news | guardian.co.uk
- Wellwishers brave beatings to visit lawyer under house arrest in China | World news | guardian.co.uk
- Ai Weiwei's mother accuses officials of hounding her son | Art and design | guardian.co.uk
- Ai Weiwei vows to clear tax charges amid fresh challenge from authorities | Art and design | guardian.co.uk
- Ai Weiwei supporters send money for tax bill – in pictures | Art and design | guardian.co.uk
- Liu Xiaobo: new book lifts China's gag on jailed Nobel peace prizewinner | World news | The Observer
- Ai Weiwei ordered to pay £1.5 million in tax | World news | guardian.co.uk
- BP's $7bn sale of Pan American Energy to Argentinian firm collapses | Business | The Guardian
- BP's bid to clean up its act dealt blow by revelations in Russia case | Business | The Observer
- Russia enthralled by oligarch heavyweight court bout - Roman Abramovich v Boris Berezovsky | World news | The Guardian
- New Zealand prime minister seeks to block recording | World news | guardian.co.uk
- Julian Assange awaits high court ruling on extradition | Media | guardian.co.uk
- Julian Assange seeks to take extradition fight to supreme court | Media | guardian.co.uk
- Julian Assange's options narrow as judges reject extradition appeal | Media | The Guardian
- UK Border Agency officials 'illegally targeting' bus passengers | UK news | The Observer
- Global campaign to decriminalise homosexuality to kick off in Belize court | World news | The Guardian
- Light aircraft crashes into the English Channel | UK news | The Observer
- Iran protesters attack UK embassy in Tehran - live | World news | guardian.co.uk
- Britain withdraws diplomats from Iran after embassy attack | World news | guardian.co.uk