

READING THE NEWS THROUGH ITS STRUCTURE: NEW HYBRID CONNECTIVITY BASED APPROACHES

Programa de Doutoramento em Ciências da Complexidade
Doctoral Programme in Complexity Sciences

Orientador / Advisor: Professor Jorge Manuel Anacleto Louçã

David Manuel de Sousa Rodrigues

March 17, 2014

Outline of presentation

- Context of this work and Related Work
 - Newspapers
 - Adaptive Networks
 - Q-analysis
 - Community detection
 - Ant Colony Optimisation
- Hybrid Connectivity Based Approaches
 - Variation of Information and Dynamic Networks
 - Clustering News: Timelines with k-means
 - Clustering News: Community finding with Q-analysis filtering
 - Hamiltonian Paths in Q-analysis eccentricity matrices
- Conclusions

Objectives

- The thesis presents four approaches to the problem of **identifying meaningful structure** in the news published online.
- This is a hard problem due to the high volume of produced data and to the possible high dimensionality of the data collected.

Contributions

- The thesis shows how Hybrid Connectivity Based Approaches give insights to news structure.
 - Adaptive Networks and Mutual Information
 - Clustering with k-means and feature vectors
 - Clustering news with pre-filtering with Q-analysis
 - Creating Hamiltonian paths of news using Q-analysis eccentricity as distances.
 - New Ant Colony Optimisation Algorithm

CONTEXT AND RELATED

Part I

Context: newspapers (print)

Portuguese circulation

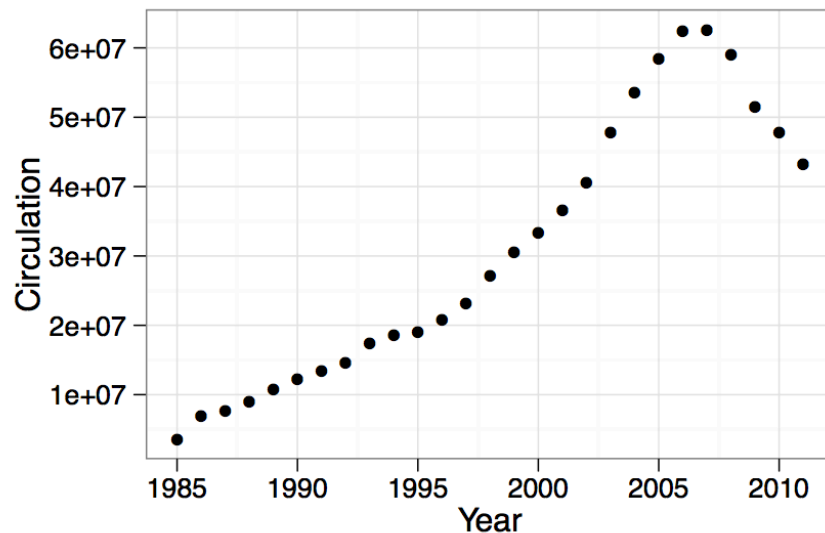


Figure 1.1: Portuguese journal+magazine printed circulation

UK Circulation

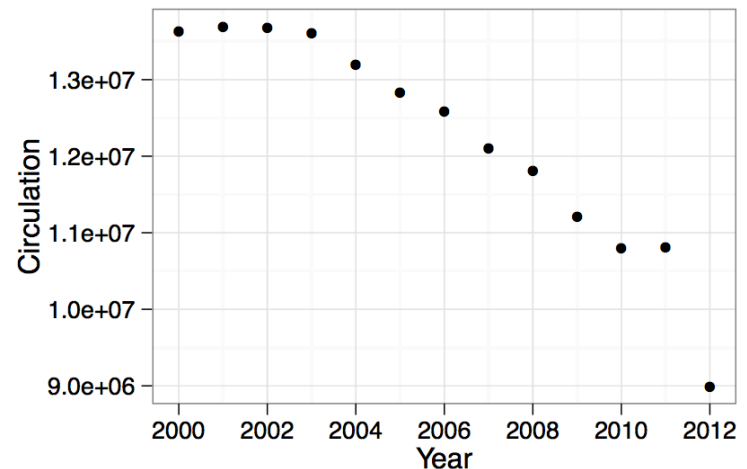


Figure 1.2: Aggregated UK 13 top journals daily printed circulation

Context: newspapers (electronic)

Internet traffic

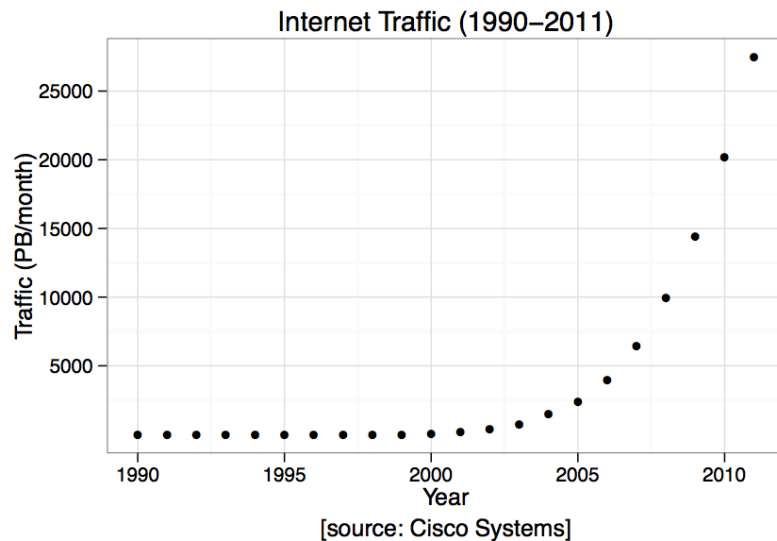


Figure 1.3: Internet Traffic in 1990-2011 (in PetaBytes/month)

Internet overtakes print as news outlet

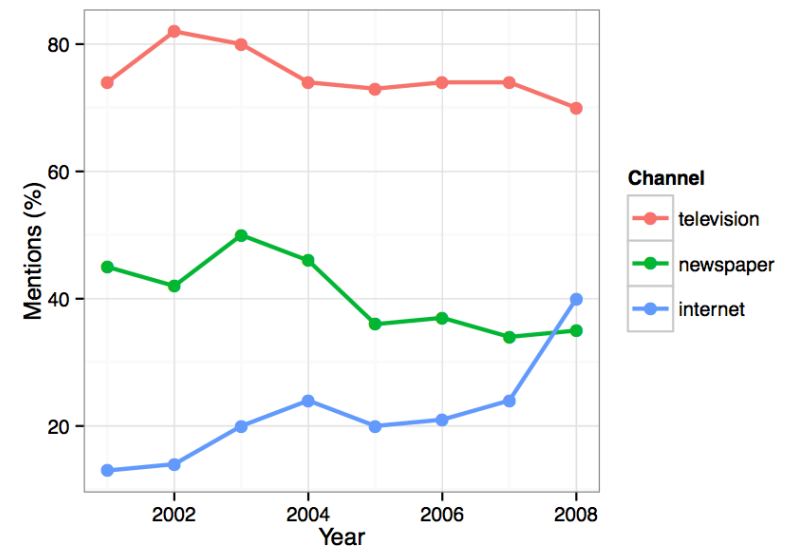


Figure 1.4: Internet overtakes newspapers as news outlet
: (Kohut and Remez, 2008)

Related work: Document analysis

- Categorisation of documents (Supervised)
 - Machine learning
 - K-neighbours, SVM, NN, etc...
- Clustering (unsupervised)
 - Document navigation
 - Sometimes associated with clustering
 - Information retrieval

Related work: Networks

- Network Science
 - Adaptive Networks
 - (interplay of topology dynamics and local dynamics of networks)
 - Community Detection in Graphs
 - Clustering nodes of graphs
 - Q-analysis
 - Topological description of the high dimensionality of structures.

Related: Bio-inspired

- Swarm Intelligence algorithms
 - Ant Systems
 - Ant Colony Optimisation
 - Travelling Salesman Problem
 - Anti-pheromone ideas
 - subtractive anti-pheromone (SAP)
 - 1 pheromone – subtracted from poor solutions
 - preferential anti-pheromone (PAP)
 - 2 pheromones but to solve bi-criterion optimisation problems

HYBRID CONNECTIVITY BASED APPROACHES

Part II

Research Opportunities

- **Finding Patterns in Data**
 - **Community Detection** and **Adaptive Networks**
 - **Q-analysis** to describe high dimensional structures
 - **Bio-inspired** heuristics to solve
- **Combining Different Techniques to produce better algorithms for existing problems.**

Hybrid Connectivity approaches

- **Hybrid?**

- This thesis proposes approaches that involve multiple techniques
Usually two techniques are used.

- **Connectivity?**

- Data is represented by entities and relations between them.
 - Binary relations (graphs)
 - n-ary relations (hypergraphs, etc..)

TOPIC MONITORING WITH VARIATION OF INFORMATION AND DYNAMIC NETWORKS

Description

Schematics of topic growth and detection via VI

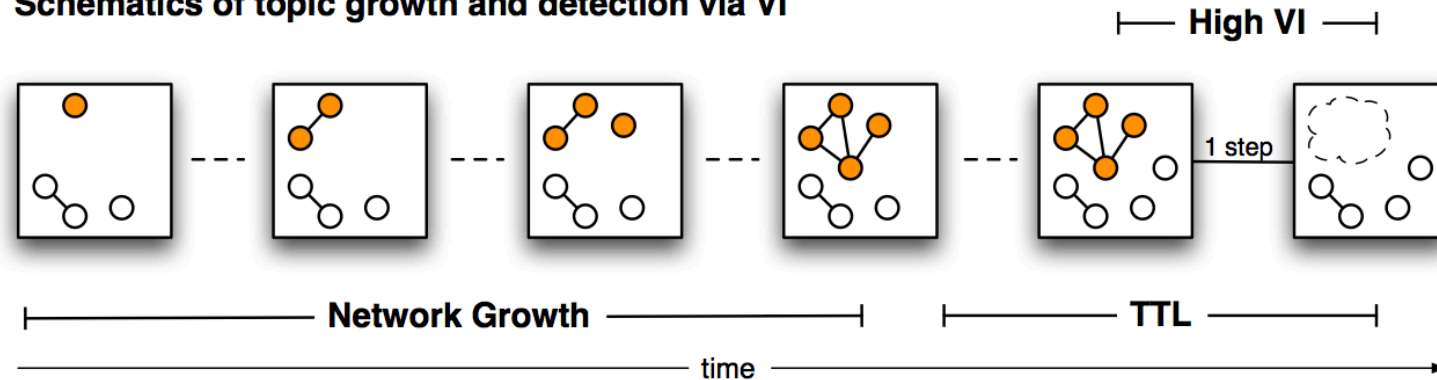


Figure 8.1: Schematics of network growth and variation of information on cluster deletion

$$VI(C, C') = H(C) + H(C') - 2I(C, C') \quad (8.5)$$

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

Main Results

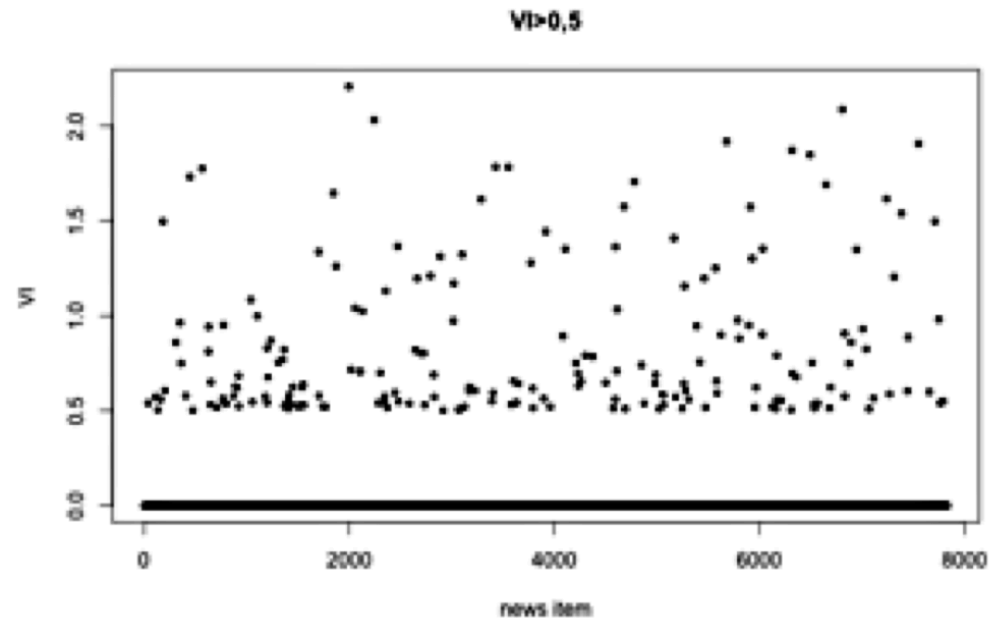


Figure 8.2: Evidence of topic deletion by tracking VI

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

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

CLUSTERING NEWS:

constructing timelines of news with *k*-means

Clustering with k-means

- Objective: create clustered timelines of news to see time-dependence of news.
 - Possibility to track back in time origins of stories
 - Create an interface for story navigation
- Approach: *tf.idf* feature vectors clustered with *k*-means
 - Write interactive software for news navigation (part of Theseus)

Clustering with k-means

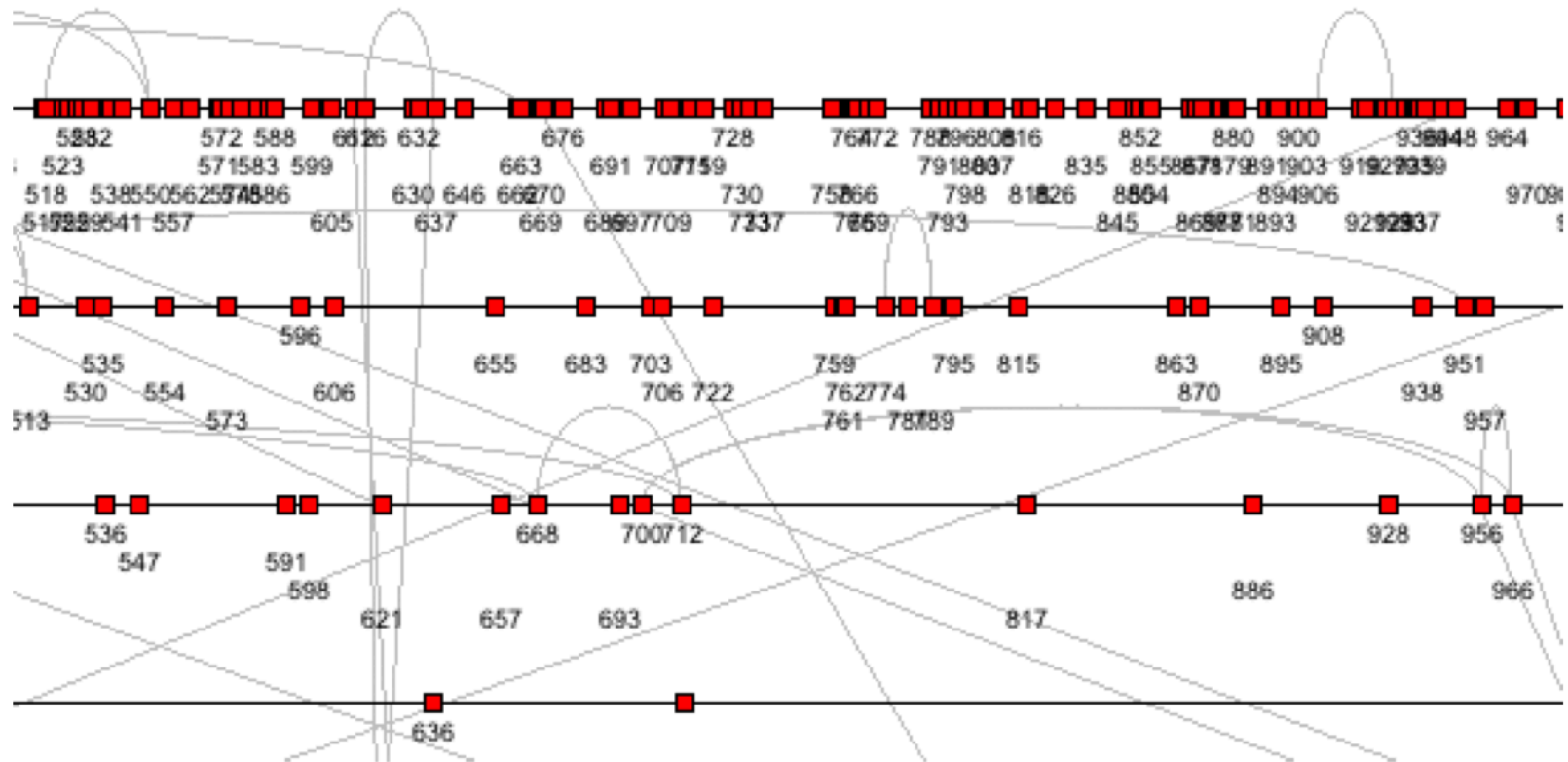


Figure 9.2: Detail of the time dependence arcs in the analysis of The Guardian timeline

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

CLUSTERING NEWS:

finding communities with Q-analysis filtering

Clustering with no filtering

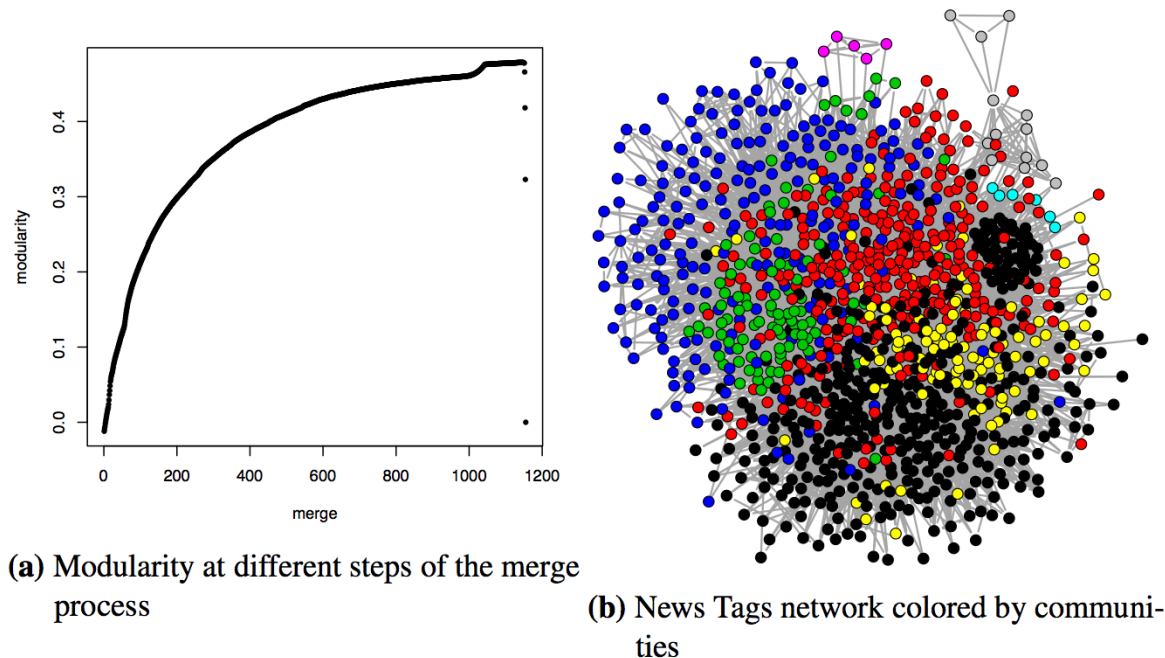


Figure 10.1: Modularity of the clustering fast greedy algorithm by (Clauset et al., 2004) and resulting communities.

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

Fraction of vertices in resulting graphs

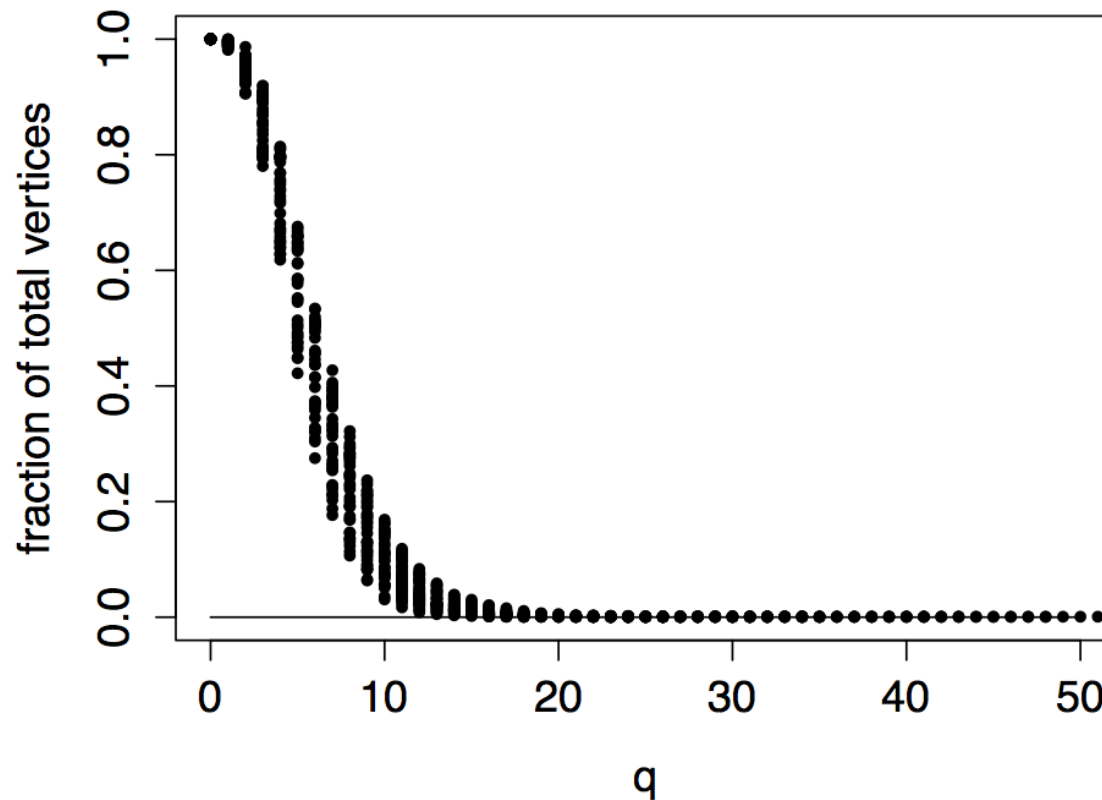


Figure 10.2: Fraction of vertices in the resulting graphs as a function of q

Fraction of vertices in maximal cluster in relation to that particular subgraph

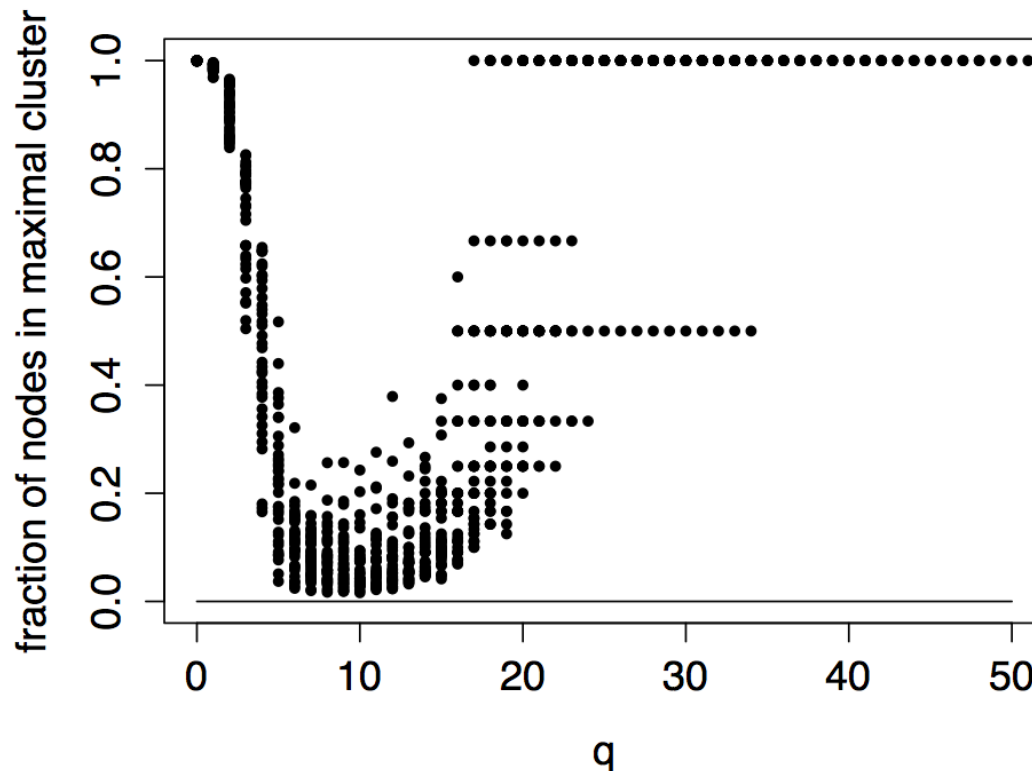


Figure 10.4: Fraction of nodes in the maximal cluster relatively to the number of nodes in that particular graph

Number of Clusters

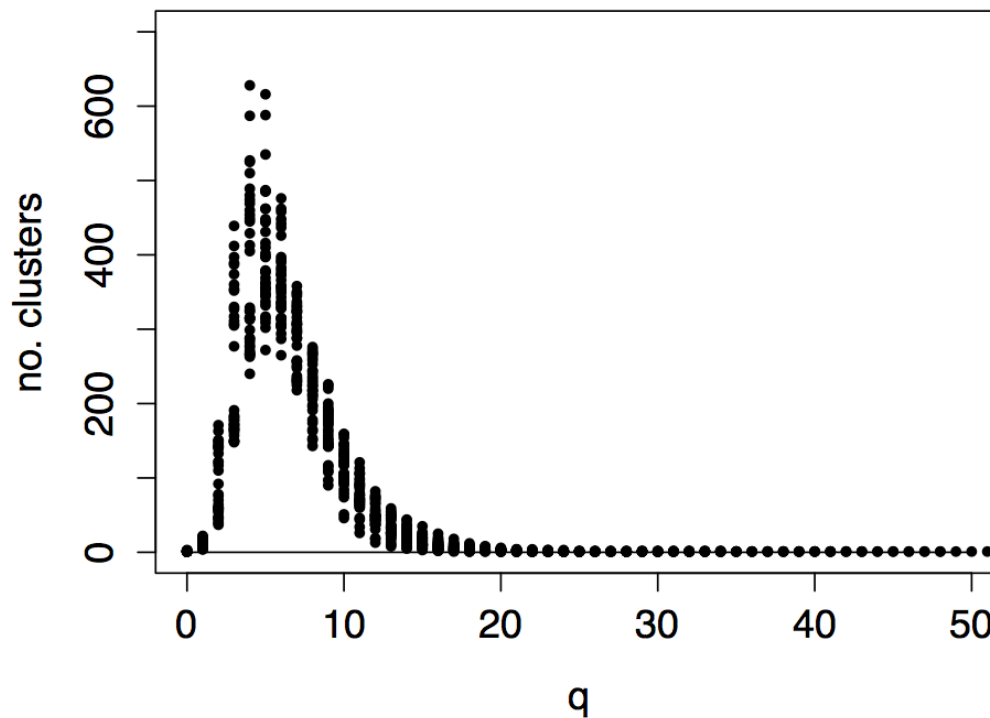


Figure 10.5: No. of clusters as a function of q

Modularity of the resulting clustering

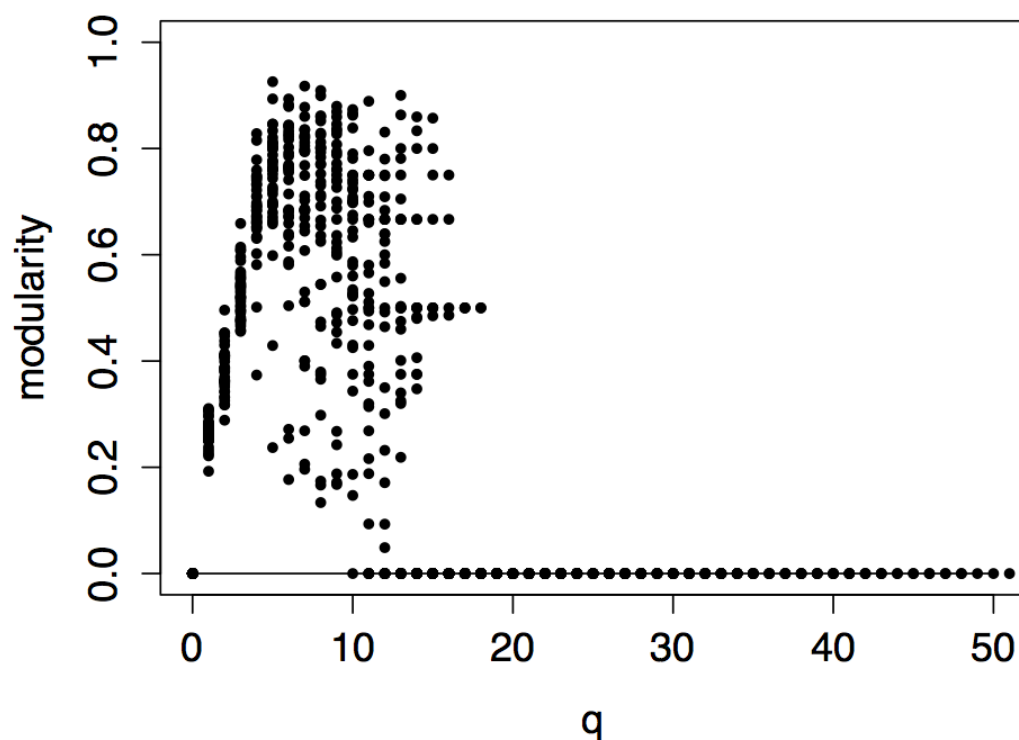


Figure 10.6: Modularity of the induced graph as a function of q

Software developed for visualisation of case study (on CD)

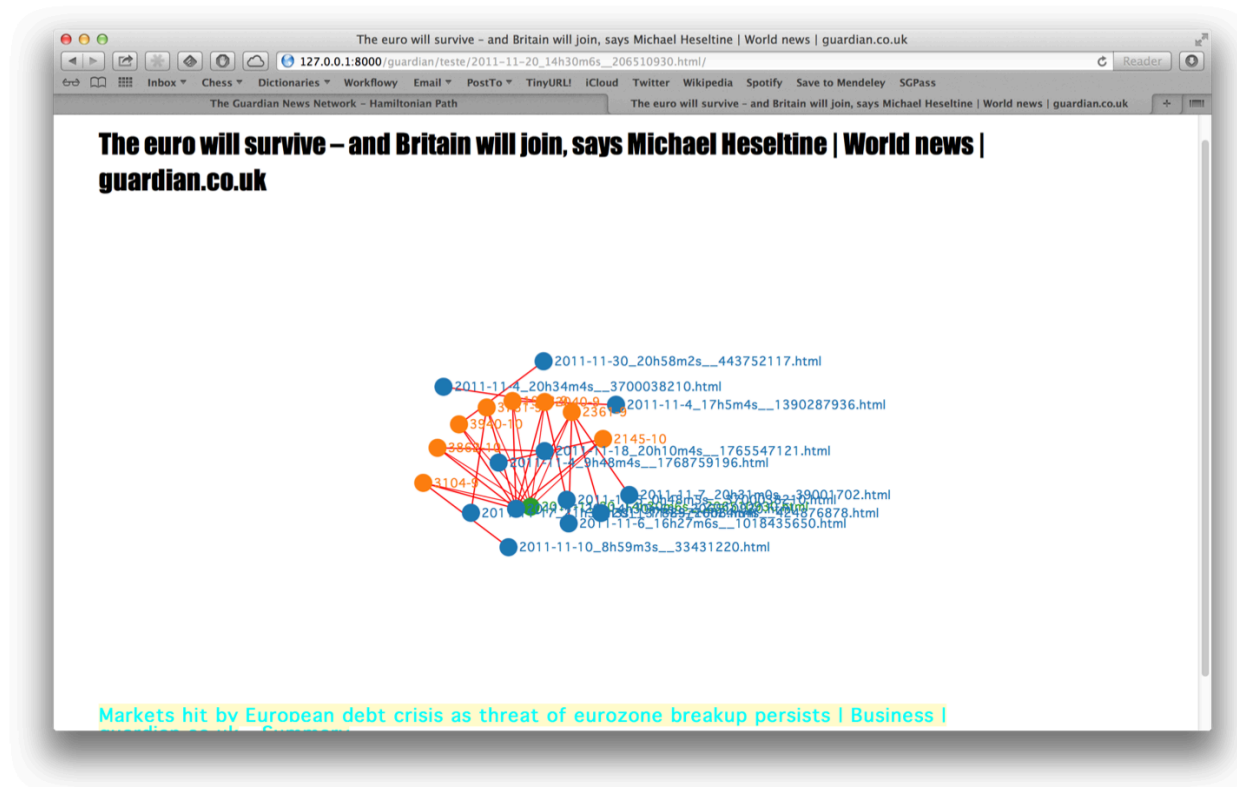


Figure 10.11: *Q*-analysis visualisation software displaying the active document (green) and the connected documents (blue) via their shared faces (orange)

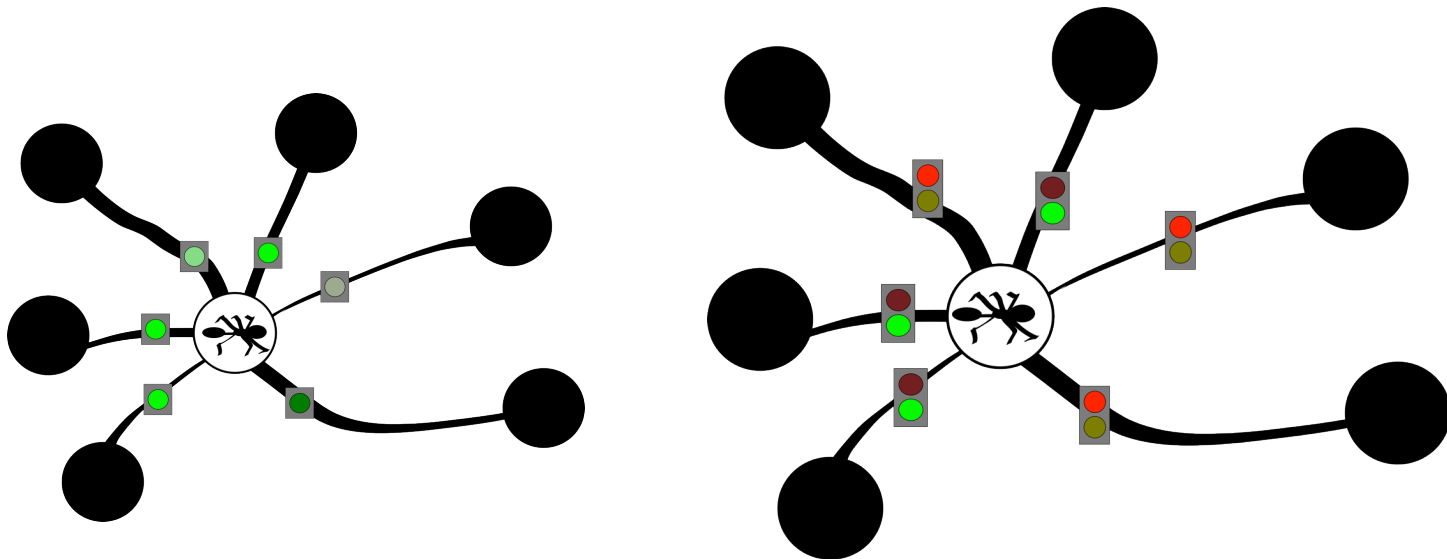
HAMILTONIAN PATHS IN Q-ANALYSIS ECCENTRICITY MATRICES

Two threads

- Development of a novel Travelling Salesman Problem algorithm
 - In collaboration with **Vitorino Ramos** [Rodrigues, 2011, Ramos 2011, Ramos 2013]
- Application of Q-analysis eccentricities matrices as distance matrices in the construction of Directed Hamiltonian Paths in the TSP problem.

2nd Order Swarm Intelligence

- Pharaoh's ants (*Monomorium pharaonis*) deposit a pheromone as a '**no entry**' signal to mark unrewarding foraging paths.
- Double Pheromone Model on top of traditional ACS.
 - Traditional positive reinforcement pheromone
 - Use of Negative Pheromone to block bad paths.



Results – Static problems

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.ifl.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

Influence of negative pheromone

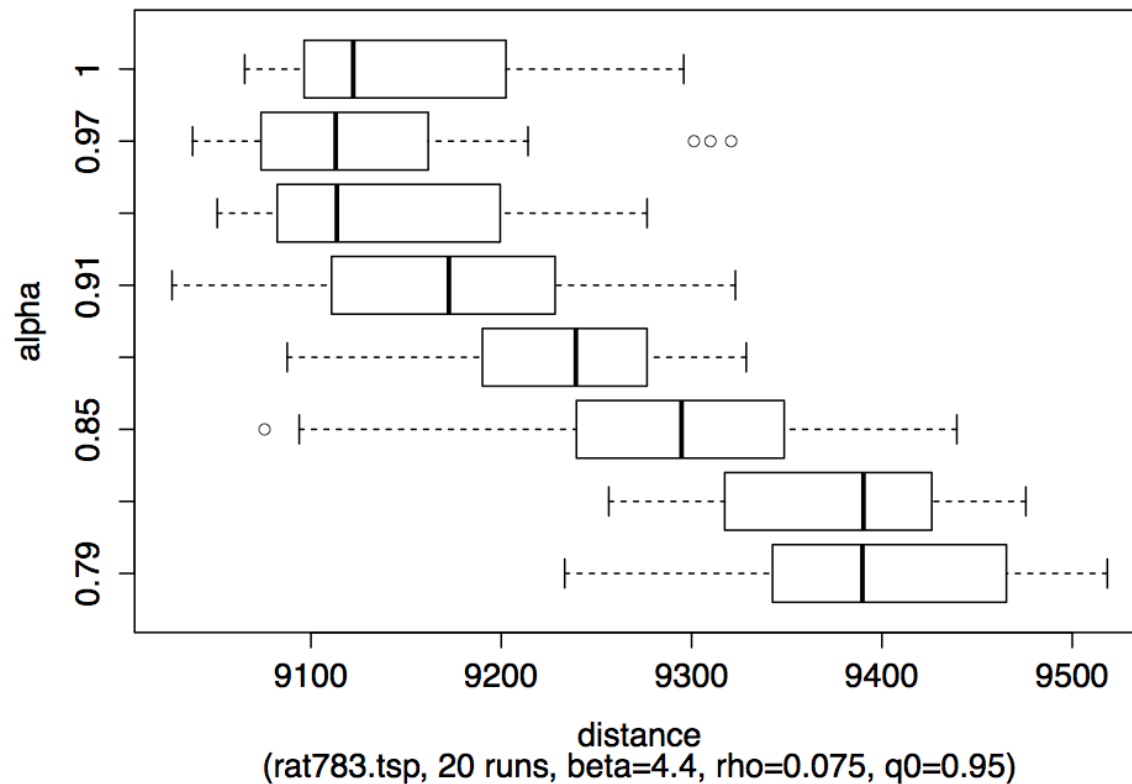


Figure 11.5: Influence of negative pheromone ($1 - \alpha$) on the TSP problem *rat783.tsp*

Application to dynamic problems: recovery patterns

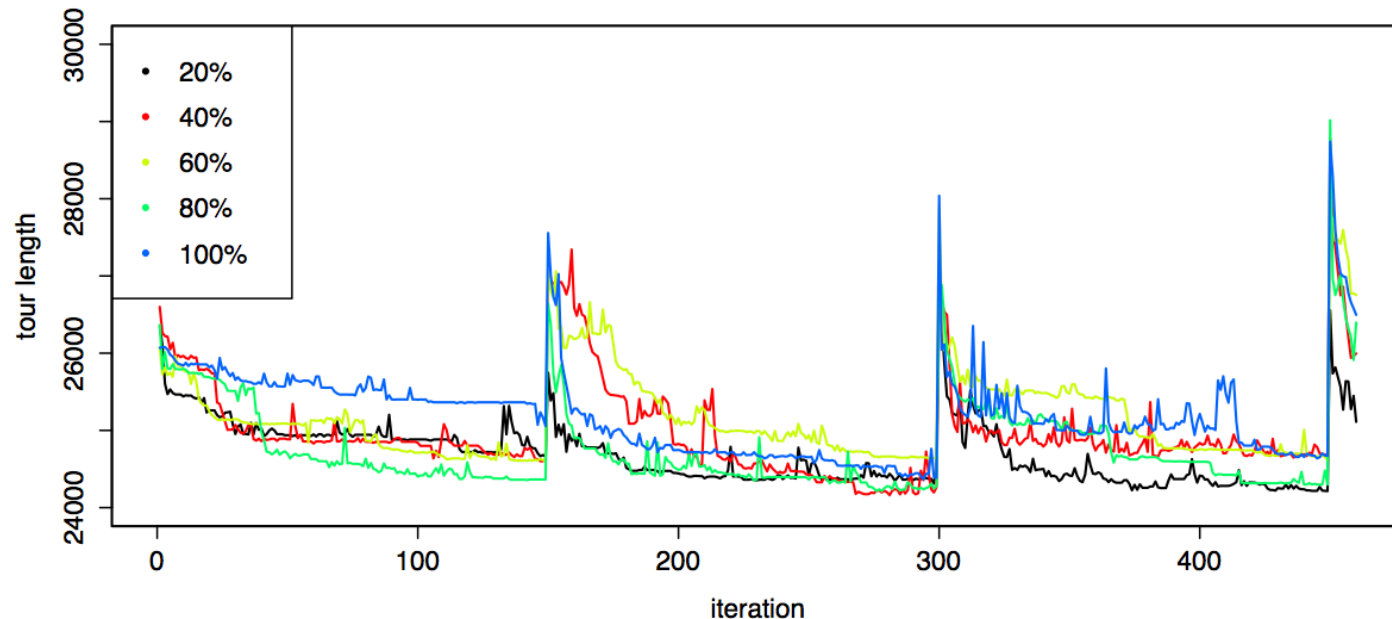


Figure 11.6: Recovery times of the Dynamical stress tests over fl1577.tsp problem (1577 nodes)
- 460 iterations - Swift changes at every 150 iterations (20%=314 nodes, 40%=630 nodes, 60%=946 nodes, 80%=1260 nodes, 100%=1576 nodes)

Application to the News

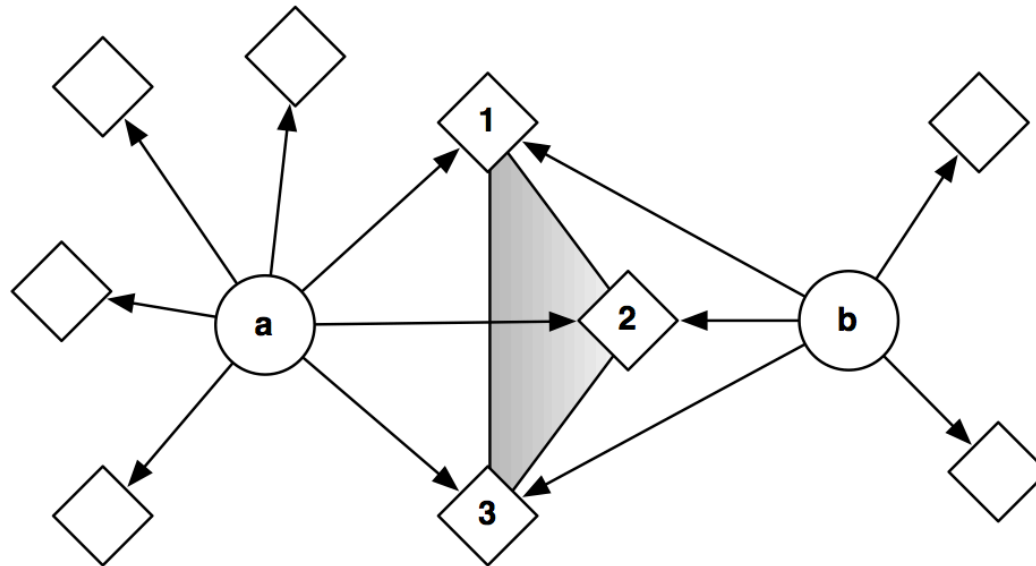


Figure 11.7: Two simplicies a and b connected by the 2-dimensional face, the triangle $\{1, 2, 3\}$.

$$ecc_{a,b} = \frac{|a| - |a \cap b|}{|a|} \quad (11.9)$$

Software Developed (on CD)

Silvio Berlusconi hints at comeback as Italy tries to form new government | World news | guardian.co.uk

[Europe](#) [European banks](#) [Italy](#) [Business](#) [Eurozone crisis](#) [World news](#) [Silvio Berlusconi](#)

Silvio Berlusconi to bow out after Italian MPs vote for savage cuts | World news | The Observer

[Europe](#) [Italy](#) [Business](#) [Eurozone crisis](#) [Global economy](#) [World news](#) [Silvio Berlusconi](#) [European Union](#)

Crucial vote for eurozone due in Italian senate | Business | The Guardian

[Europe](#) [Italy](#) [Business](#) [Eurozone crisis](#) [World news](#) [Silvio Berlusconi](#) [European Union](#)

Silvio Berlusconi to resign after austerity vote | World news | guardian.co.uk

[Europe](#) [European banks](#) [Italy](#) [Business](#) [Eurozone crisis](#) [World news](#) [Silvio Berlusconi](#)

Italy's borrowing costs keep on rising despite Berlusconi's promise to quit | Business | guardian.co.uk

[Italy](#) [Business](#) [Bonds](#) [Eurozone crisis](#) [World news](#) [Silvio Berlusconi](#)

Silvio Berlusconi vows to resign as Italy's prime minister | World news | The Guardian

[Europe](#) [Financial crisis](#) [Italy](#) [Business](#) [Eurozone crisis](#) [World news](#) [Global recession](#) [Silvio Berlusconi](#)

European debt crisis live: Greece locked in coalition talks | Business | guardian.co.uk

[Financial crisis](#) [Eurozone crisis](#) [Italy](#) [Business](#) [Greece](#)

Eurozone crisis: Spain's election leaves markets on edge | Business | guardian.co.uk

[Italy](#) [Business](#) [Eurozone crisis](#) [Market turmoil](#) [US economy](#) [Greece](#) [Spain](#)

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

CONCLUSIONS

Main Contributions of this work

- 4 approaches based on the connectivity of the system that reveal the underlying structure of the news.
- Each as advantages and disadvantages

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 Time-lines with k-means	Connectivity given directly from word frequency. <i>k</i> -means easy to implement	<i>k</i> -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.
<i>Q</i> -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 <i>q</i> . Modularity has resolution limits.
<i>Q</i> -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.

Main Contributions of this work

- 4 approaches based on the connectivity of the system that reveal the underlying structure of the news.
- Each as advantages and disadvantages
- New Optimisation bio-inspired algorithm for TSP problems (adaptable to new problems)
- Software for dealing with gathering, processing, and visualising these systems (Theseus)

Limitations and Perspectives

- *a posteriori* analysis
 - AN+VI
- user defined parameters to fit the models
 - q in Q-analysis filtering, k in k -means, TTL in AN+VI
- Size of data matrices
 - feature vectors in k -means or eccentricity matrices in SOSI.
- Information theory based measures as signal detectors for change
- Bio inspired methods for new swarm intelligent algorithms
- Topology based methods to reduce space of exploration of solutions

Limitations and Perspectives

- There is no universal solution or general panacea applicable to complex systems
- Hybrid approaches have both advantages and disadvantages.
- Complexity practitioners need to do engineer problem-driven solutions.
 - Q-analysis and any other low level description of data that manipulates data to the least are important.
 - Bio-inspired algorithms are useful
 - Traditional combinatorial algorithm will cope badly with exponential growth of data.