# Identifying news clusters using $Q$-analysis and Modularity

David Rodrigues*

**Abstract**

With online publication and social media taking the main role in dissemination of news, and with the decline of traditional printed media, it has become necessary to devise ways to automatically extract meaningful information from the plethora of sources available and to make that information readily available to interested parties. In this paper we present a method of automated analysis of the underlying structure of online newspapers based on $Q$-analysis and modularity. We show how the combination of the two strategies allows for the identification of well defined news clusters that are free of noise (unrelated stories) and provide automated clustering of information on trending topics on news published online.

## 1 Introduction

Every day Internet presents a huge amount of new information, either in personal or institutional web pages, posted in social networks or as a reflex of the media (TV, newspapers) regular diffusion of news. Regarding this last case, information flows on media as a result of complex relations between facts and news, shaping a ever-changing network of relations between topics. Topic detection and the unveiling of dynamic relations between topics can give new insights to the understanding of communication mechanisms in human societies.

Recent research in the domain of topic detection applied different techniques, including regression models, nearest neighbor classification, Bayesian probabilistic approaches, decision trees, inductive rule learning, neural networks, online learning and Support Vector Machines (Cardoso-Cachopo and Oliveira 2003; Yang and Liu 1999; Miao and Qiu 2009; Joachims 1998; Hamamoto et al. 2005; Solé et al. 2010). However, most of those approaches 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. These approaches face two major restrictions: they are language-dependent, requiring the work of specialists for analysing and classifying; and they are not adapted for finding new categories in data without re-training.

Attempts to solve the above problem include the use of using a dynamic network where a time sliding window is used and changes between consecutive generated networks are evaluated for the variation of information between consecutive windows (Meilă 2007; Rodrigues 2010).

Also, previous community detection algorithms in graphs aimed at the inclusion of all nodes of the graph into a cluster. This process was traditionally considered exclusive and a node that belong to some cluster wouldn't belong to other cluster. To solve this several algorithms have been proposed that allows the overlapping of communities. One example of such algorithms is the clique percolation method by Derenyi, Palla, and Vicsek 2005; Palla et al. 2005. On the other hand of the spectrum, very little attention has been given to to the cases where not all nodes should be included in a cluster. This nodes nodes aren't really associated with any other node in a meaningful way and traditional clustering algorithms don't have mechanisms to deal with them. The main reason for this is that they lack the structural

*david.rodrigues@open.ac.uk

information needed to include or reject them from the clustering process. Each edge is unidimensional representing a binary relation and the only additions to this simplification are the inclusion of directionality and weights to these relations. We believe that on the other hand, complex systems being composed of *n*-ary relations, should be described in languages that support these high dimensional relations. By including structural information on the connectivity of the graph one can filter out those nodes that otherwise would be misclassified by traditional algorithms. For this we use *Q*-analysis (Johnson 1970; Atkin 1972, 1974; Beaumont and Gatrell 1982; Johnson 1983) and Modularity based clustering (Clauset, Newman, and Moore 2004; Newman 2006).

In this work we propose a method for the automatic classification of newspaper news based on the simplicial complex generated from the news published online and from the tags associated with those news entered by journalists at publication time. We analyse this system by first using *Q*-analysis on the simplicial complex and then by clustering the higher *q*-connectivity induced subgraphs. After extraction of this high connectivity graph one can proceed to cluster the news stories by using any traditional graph clustering algorithm (for a complete review of community detection algorithms and strategies see Fortunato 2010). In this case we show the results of applying a modularity optimization (Clauset, Newman, and Moore 2004) based method to the analysis of the news published online by *The Guardian*[1] and how the *Q*-analysis improves the quality of the clustering.

## 2 Discussion of Results

*The Guardian* classifies every news 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 characterize that particular story with. We take advantage of this human labelling to characterize the structure of the network created by all the news of *The Guardian*.

Initially a graph is constructed by defining each published story as a node in the graph and then by considering that two stories were connected if they shared at least one tag among them. We can think of this approach as the construction of the graph from the *0*-connected simplicial complex obtained in *Q*-analysis. This corresponds to the simplest projection with the broadest structural information.

The application of community detection algorithms based on modularity optimization generates a clustering where the maximal value of modularity is $0.48$ for a total of $9$ communities.

**Table 1:** Cluster sizes for *The Guardian* news published during the month of November, 2011

| id: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| size: | 363 | 303 | 96 | 221 | 6 | 5 | 102 | 13 | 46 |

This clustering reveals large components meaning that probably a division into more components would be of great interest and that these large clusters don't capture the fine structure of the news, due to the resolution limit of modularity optimization methods (Fortunato and Barthelemy 2006). For this we need a new approach, one that effectively can give insights into the news structure and that removes noise news [2] that otherwise would be part in spurious classifications.

The previous analysis of the Tag network of *The Guardian* news shows clearly that calculating the clusters via modularity optimization alone isn't enough. In it we considered full connectivity between

[1] The Guardian (`http://www.guardian.co.uk/`)

[2] news that aren't highly connected to others and therefore can be discarded when considering the task of understanding the main themes of a newspaper

nodes of the graph (two nodes where connected if they shared at lease one tag). We now show that by extracting the induced subgraph of higher connected nodes we can obtain a clearer separation of modules in the news structure.

For this we proceeded with a $Q$-analysis of the news-tags system. In $Q$-analysis (Johnson 1981, 2005, 2006) two nodes are connected by a link if they share at least $(q + 1)$ common attributes. In this case two news are connected if they share at least $q + 1$ tags. The resulting networks constructed in this way guarantee a minimum level of connectedness between published news.

We characterized the induced subgraphs in terms of several properties (following figures) as a function of $Q$.

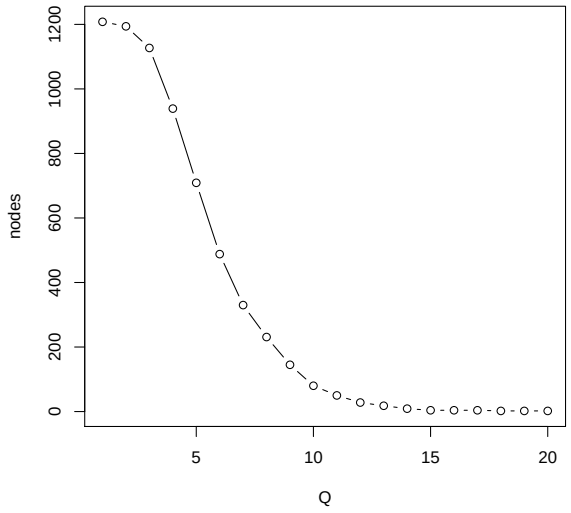


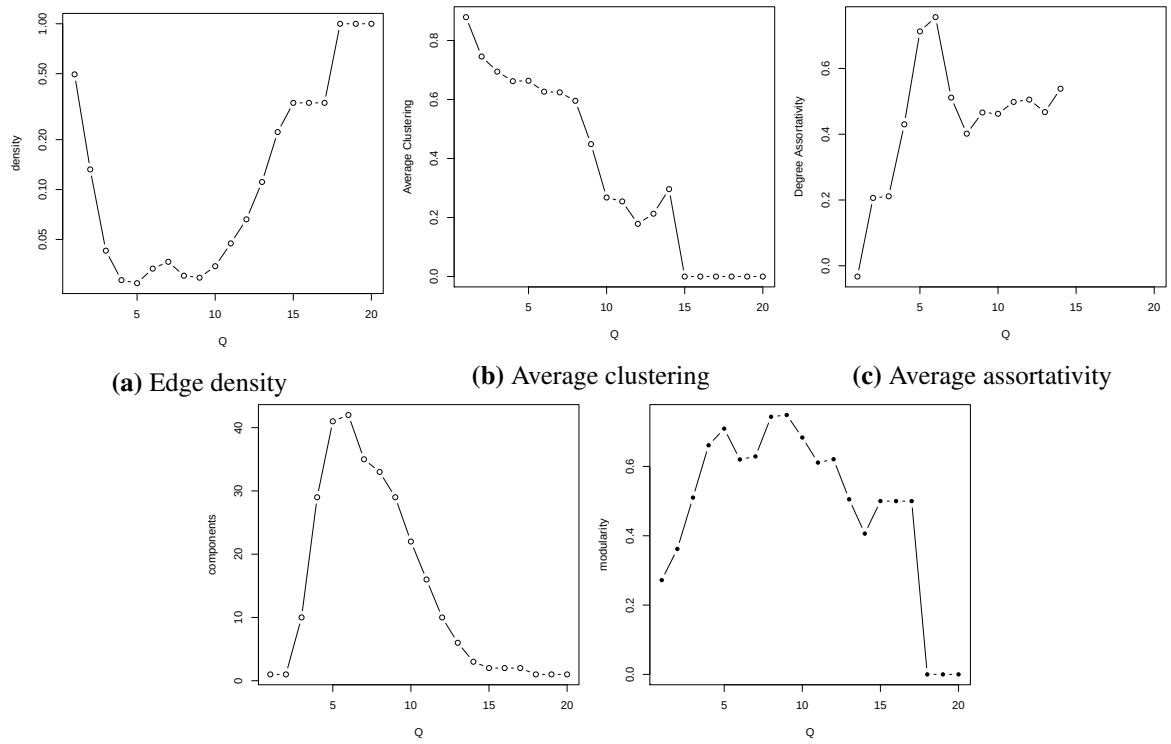

**Figure 1:** Number of nodes of the induced subgraph as function of $Q$

From figure 1 it is clear that the number of nodes present in the induced subgraph drops abruptly for higher values of connectivity.

From figure 2 we show how edge density, average clustering, degree assortativity and number of components and modularity change with the increased connectivity of the induced subgraphs.

It is clear that to identify the most significant components in the news published one is interested in some useful characteristics of the induced graph. It should have high clustering, high number of components, and also high degree assortativity. It is expected also that the value o modularity for this graph is also high. Usually one wants to choose an induced graph such that the number of nodes present is still high. In the case of the Guardian we can see that by choosing $Q = 5$ we can have these characteristics.

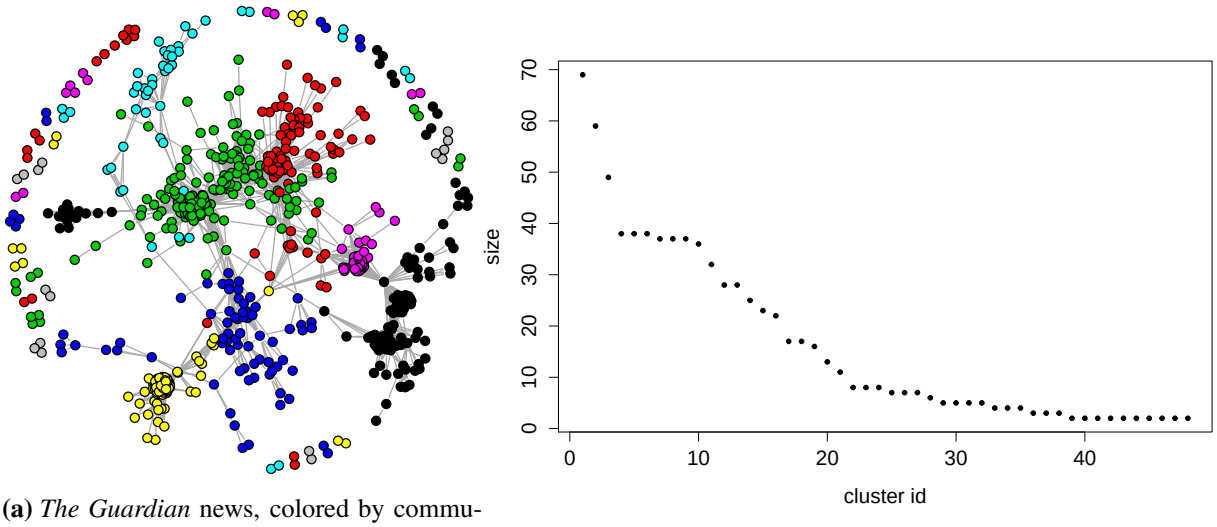
As we can see in figures 2, 3a and 3b, the clustering process case gives higher values for modularity of the resulting induced graphs. This gives high confidence on the structural properties of the clusters identified. Manual inspection of the clusters was conducted and revealed a total absence of unrelated stories in the identified clusters. By using this system the we were able to identify $48$ clusters instead of the previous $9$ with higher modularity. This is an indication that the previous modularity optimization had limitations (mainly because of the noise stories) and it was merging into the same cluster unrelated news stories. The size distribution shows the presence of many small sized cluster (typically between 2 and 5 news) but at least 10 clusters are constituted by 30 or more news stories. These are the main topics being discussed during that month and include the scandal of James Murdoch news empire, the evolution of the Syrian situation and the Arab Spring and the Eurozone debt crisis, among others.




**(a)** Edge density **(b)** Average clustering **(c)** Average assortativity

**(d)** Number of components **(e)** Modularity

**Figure 2:** Analysis of the induced subgraphs for the *The Guardian*




**(a)** *The Guardian* news, colored by community: $Q = 5$, $Modularity = 0.71$

**(b)** Size distribution of clusters of *The Guardian* news (Q=5)

**Figure 3:** Clusters in *The Guardian* news during November 2011

# 3 Conclusions

In this paper we present a novel way for filtering news stories published in online newspapers by using $Q$-analysis and modularity optimization. The process deals with the problem of selecting which relevant stories should be considered when clustering documents. We show how a bipartite graph formed from the documents and the tags associated with each document can be used to filter the stories that aren't strongly connected to other members of the graph.

The main advantage of using this technique is that the process of analysis is automated. On the other hand the process poses the disadvantage of discarding those nodes that aren't highly connected, but this disadvantage is a parameterized one and can be minimized by lowering the value of connectedness needed for inclusion in the induced subgraph. If all nodes are to be included then the problem reverts to a traditional clustering problem.

This method presents itself to many applications where the structural properties of the system are of high importance. Traditional clustering algorithms traditional consider the clustering process on graphs where nodes are equivalent and where the different structural connectivities of the nodes isn't taken into account. By using extra information from the bipartite graph, through $Q$-analysis, it is possible to filter relevant information from spurious information in an automated way.

In the case of the corpus presented here, it relies on human labelling of the documents, but the construction of the bipartite graph can be done with one of the many automated topic modelling strategies developed in recent years (see Blei, Ng, and Jordan 2003; Li and McCallum 2006).

Another advantage of this method is that it can be applied to different corpora of textual based documents. Instead of analysing a single newspaper, it can be applied to daily news from different newspapers or channels to extract relevant stories and topics. It provides a way of clustering relevant trends in the news almost in real time. The method can also be applied to other fields: automated clustering of new scientific publishing, tracking of opinion dynamics on social media, and detection of brand awareness in online communities, constitute potential applications where knowledge of the structural properties of the system improves the quality of the analysis.

# References


Atkin, Ronald Harry (1972). "From cohomology in physics to q-connectivity in social science". In: *International Journal of Man-Machine Studies* 4.2, pp. 139 –167. ISSN: 0020-7373. DOI: 10.1016/S0020-7373(72)80029-4.

— (1974). *Mathematical Structure in Human Affairs*. 1st ed. 48 Charles Street, London: Heinemann Educational Publishers.

Beaumont, John R and Anthony C Gatrell (1982). *An introduction to Q-analysis*. Norwich Norfolk: Geo Abstracts. ISBN: 086094106X.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (Mar. 2003). "Latent dirichlet allocation". In: *J. Mach. Learn. Res.* 3, pp. 993–1022. ISSN: 1532-4435.

Cardoso-Cachopo, Ana and Arlindo L Oliveira (2003). "An Empirical Comparison of Text Categorization Methods". In: *String Processing and Information Retrieval*. Ed. by Mario A Nascimento, Edleno S De Moura, and Arlindo LEditors Oliveira. Springer Verlag, Heidelberg, DE, pp. 183–196.

Clauset, Aaron, M. Newman, and Cristopher Moore (2004). "Finding community structure in very large networks". In: *Phys. Rev. E* 70 (6). Phys. Rev. E 70, 066111 (2004), p. 066111. DOI: 10.1103/PhysRevE.70.066111.

Derenyi, Imre, Gergely Palla, and Tamas Vicsek (2005). "Clique percolation in random networks". In: *Physical Review Letters* 94, p. 160202.

Fortunato, Santo (2010). "Community detection in graphs". In: *Physics Reports* 486.3-5, pp. 75–174.
Fortunato, Santo and Marc Barthelemy (July 2006). "Resolution limit in community detection". In: *physics/0607100*. Proc. Natl. Acad. Sci. USA 104 (1), 36-41 (2007). DOI: doi:10.1073/pnas.0605965104.
Hamamoto, M. et al. (Apr. 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*. IEEE, pp. 122–127. ISBN: 0-7695-2414-1. DOI: 10.1109/WIRI.2005.1.
Joachims, Thorsten (1998). "Text categorization with support vector machines: Learning with many relevant features". In: *Machine Learning ECML98* 1398.23. Ed. by Claire Nédellec and CélineEditors Rouveirol, pp. 137–142.
Johnson, J. H. (Feb. 1970). "Q Analysis of Large Samples". In: *Journal of Marketing Research* 7.1. ArticleType: research-article / Full publication date: Feb., 1970 / Copyright © 1970 American Marketing Association, pp. 104–105. ISSN: 0022-2437. DOI: 10.2307/3149515.
— (1981). "Some structures and notation of Q-analysis". In: *Environment And Planning B* 8, pp. 73–86.
— (Sept. 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*.
— (2005). "Multidimensional Multilevel Networks in the Science of the Design of Complex Systems". In: *ECCS 2005 Satellite Workshop: Embracing Complexity in Design*. Ed. by Jeffrey Johnson. Vol. ECCS 2005 Satellite Workshop: Embracing Complexity in Design.
— (2006). "Can Complexity Help Us Better Understand Risk?" In: *Risk Managment* 8.4, pp. 227–267. ISSN: 1460-3799.
Li, Wei and Andrew McCallum (2006). "Pachinko allocation: DAG-structured mixture models of topic correlations". In: *Proceedings of the 23rd international conference on Machine learning*. ICML '06. Pittsburgh, Pennsylvania: ACM, pp. 577–584. ISBN: 1-59593-383-2. DOI: 10.1145/1143844.1143917.
Meilă, Marina (2007). "Comparing clusterings—an information based distance". In: *J. Multivar. Anal.* 98.5, pp. 873–895.
Miao, Youdong and Xipeng Qiu (2009). "Hierarchical Centroid-based Classifier for Large Scale Text Classification". In: *Large Scale Hierarchical Text classification (LSHTC) Pascal Challenge*.
Newman, M. (2006). "Modularity and community structure in networks". In: *Proceedings of the National Academy of Sciences* 103.23, pp. 8577–8582. DOI: 10.1073/pnas.0601602103. eprint: http://www.pnas.org/cgi/reprint/103/23/8577.pdf.
Palla, Gergely et al. (June 2005). "Uncovering the overlapping community structure of complex networks in nature and society". In: *Nature* 435. PMID: 15944704, pp. 814–8. ISSN: 1476-4687. DOI: nature03607.
Rodrigues, David M. S. (Sept. 2010). "The Observatorium – The structure of news: topic monitoring in online media with mutual information". In: *Proceddings of the European Conference on Complex Systems*. Ed. by Jorge Louçã. Complex Systems Society.
Solé, Ricard V. et al. (2010). "Language networks: Their structure, function, and evolution". In: *Complexity* 15.6, pp. 20–26. ISSN: 1099-0526. DOI: 10.1002/cplx.20305.
Yang, Yiming and Xin Liu (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. Berkeley, California, United States: ACM, pp. 42–49. ISBN: 1-58113-096-1. DOI: http://doi.acm.org/10.1145/312624.312647.