The community structure of a multidimensional network of news clips

José Luís Devezas* and Álvaro Reis Figueira

CRACS/INESC TEC,
Faculdade de Ciências,
Universidade do Porto,
Rua do Campo Alegre,
1021/1055, 4169-007 Porto, Portugal
E-mail: jld@dcc.fc.up.pt
E-mail: arf@dcc.fc.up.pt
*Corresponding author

Abstract: We analysed the community structure of a network of news clips where relationships were established by the co-reference of entities in pairs of clips. Community detection was applied to a unidimensional version of the news clips network, as well as to a multidimensional version where dimensions were defined based on three different classes of entities: places, people, and dates. The goal was to study the impact on the quality of the identified community structure when using multiple dimensions to model the network. We did a two-fold evaluation, first based on the modularity metric and then based on human input regarding community semantics. We verified that the assessments of the evaluators differed from the results provided by the modularity metric, pointing towards the relevance of the utility and network integration phases in the identification of semantically cohesive groups of news clips.

Keywords: community structure; multidimensional networks; co-reference networks; news clips.

Reference to this paper should be made as follows: Devezas, J.L. and Figueira, Á.R. (2013) 'The community structure of a multidimensional network of news clips', *Int. J. Web Based Communities*, Vol. 9, No. 3, pp.411–429.

Biographical notes: José Luís Devezas received his MSc in Informatics and Computing Engineering in 2010 from the Faculty of Engineering of the University of Porto. He is working on his PhD application at the Faculty of Sciences of the University of Porto. Currently, he is a researcher for the Breadcrumbs project at the Center for Research in Advanced Computing Systems (CRACS), which is affiliated with INESC TEC, an associate laboratory for technology and science. His research interests include network science, information retrieval and data mining, with a special interest in complex networks and its community structure.

Álvaro Reis Figueira graduated in Mathematics Applied to Computer Science in the Faculty of Sciences (UP) in 1995. He received his MSc in Foundations of Advanced Information Technology from the Imperial College, London, in 1997, and PhD in Computer Science from UP, in 2004. Presently, he is a Lecturer at the Faculty of Sciences and has been interested in text mining and information retrieval. His research interests also include e-learning, web-based learning and standards in education. His recent work involves integrating automatic classification of texts with social classification, in the scope of digital libraries and cyber-journalism.

1 Introduction

In the analysis of real-world networks, community detection methodologies enable the identification of latent semantic groups. Learning this hidden structural feature of a network is an essential step in the knowledge discovery process (e.g., the community structure of a KeyGraph (Ohsawa and Benson, 1998) can be used for event detection (Sayyadi et al., 2009), useful in disaster tracking for humanitarian decision-making (Edmonds et al., 2010), or in the identification, tracking and error preventing in medical procedures, based on the electronic medical records of the patients (Kushima et al., 2012). As methodologies evolved (Fortunato, 2010; Xie et al., 2011), scientists progressively improved their algorithms to better reflect real-world problems (Doreian et al., 2004a; Palla et al., 2005; Mucha et al., 2010; Leskovec et al., 2008; Jin et al., 2011). They first developed heuristics to identify the community structure of unidimensional networks, but quickly extended this to multidimensional and multimodal networks, as these can sometimes better reflect reality, where multiple types of relationships and nodes can coexist as indicators of a grouping behaviour. On the other hand, nodes in real-world networks often belong to more than one group, so the initial disjoint methods of partitioning a network evolved into overlapping community detection methodologies, introducing fuzziness to node membership. Motivated by the constant adaptation of community detection methodologies to real-world scenarios and aware of the growing complexity (Neubauer and Obermayer, 2009; Gargi et al., 2011; Leung et al., 2008; Leskovec et al., 2010; Šubelj et al., 2011) that they bring into the process of group identification, we studied the community structure of a multidimensional network of news clips, where connections represented co-references to either a person, a place, or a date. Our goal was to understand the advantages of considering three independent edge dimensions, as opposed to the traditional approach of establishing connections without an edge type, while simultaneously evaluating the quality of the community structure of a network created from the relationships of documents in a digital library.

We used a pragmatic approach based on data provided by the 'Breadcrumbs' system (Figueira et al., 2009). Breadcrumbs is a social network based on the relations established by collections of text fragments taken from online news. It uses social web tools to gather the opinions of readers, and creates a semantically organised model of the readers' opinions. In particular, Breadcrumbs focuses on collecting news fragments from online news sources, organising those fragments automatically in a personal digital library (PDL), and aggregating the fragments across readers and different PDLs. As part of the system, it is important to answer three of the six most important questions for (online) journalism: Who? Where? When? (The other questions are: What? How? And Why?) Accordingly, we identified expressions that conveyed the semantics for the identification of a person/people, places/locations and dates/time periods. From this new data, we established links between the news clips based on co-occurrence (Devezas et al., 2012). Each of the discovered links provided a different semantic bind that might be later on explored by a reader or a journalist. In any case, and from a global point of view, this set of links enabled the creation of a multidimensional network of news clips.

In this study, we analysed the grouping behaviour of entities by taking advantage of the diversity of latent connections in our data. The main contribution of our work is the comparison of the impact of multidimensional and unidimensional network models on the quality of the community structure.

2 Reference work

We present an overview of some of the community detection methodologies relevant to this work, explaining their individual limitations in different real-world network models. We then describe a unified view and integration methodology to extend the implementation of some of the traditional community detection algorithms to the task of detecting communities in multidimensional networks, presenting an example based on a simple artificial network.

2.1 An overview on community detection methodologies

Most community detection algorithms have a set of preconditions regarding the model and topology used to describe the network. Even though some of these preconditions can be overcome – for example, if the precondition was that the graph should be connected, communities could be identified for each separate connected component –, other preconditions must be satisfied – for example, a community detection method that only works for undirected graphs, cannot be used in directed graphs unless they previously suffer a conversion process. Losing edge direction would certainly allow the identification of the community structure using such algorithms, but that information would be discarded and have no influence whatsoever in the results. The same happens when our data can best be described using a multidimensional network model: the network can be converted to conform with the algorithms restrictions, but the additional information provided by the model is lost when extracting the latent community structure. Algorithms for community detection are evolving with this into account, which leaves room to question how beneficial it really is to use the additional information in multidimensional networks models.

We study this by using the modularity maximisation method by Newman (2006a, 2006b) to identify the community structure of a news clips network, both using a unidimensional model and a multidimensional model to describe the connections in the network. Across this paper, we often refer to this method as the Newman's method or Newman's leading eigenvector community detection methodology. This algorithm is based on the optimisation of a quality metric called modularity (Newman and Girvan, 2004), that quantifies how good a partition for a particular network is. The optimisation of the modularity is done by using the modularity matrix B:

$$B_{ij}^{(i)} = A_{ij}^{(i)} - \frac{k_i k_j}{2m}$$

where A is the adjacency matrix, k_i and k_j are the degrees of nodes i and j, and m is the total number of edges. Then, the leading eigenvectors of B are aggregated in a matrix, that is then clustered using the k-means algorithm. This yields a partition matrix containing the identified communities.

However, using this methodology by itself will not capture the additional information provided by the different dimensions in a multidimensional network. In the next subsection, we describe a methodology by Tang et al. (2011) that takes into account the common characteristics of several community detection algorithms in order to make multidimensional community detection possible.

2.2 The Tang methodology

Our study is based on the unified view and integration strategies proposed by Tang et al. (2011). They analysed several traditional community detection methodologies – latent space models (Borg and Groenen, 2003), block model approximation (Doreian et al., 2004b), spectral clustering (von Luxburg, 2007), and modularity maximisation (Newman, 2006a, 2006b) – and proposed a unified view, by identifying common steps taken by the four approaches. These steps are directly related to the four types of matrices used in unidimensional community detection:

1 The adjacency matrix, representing the network.

2 The utility matrix, which is specific to each algorithm.

3 The structural features matrix, which aggregates the eigenvectors with highest or lowest eigenvalues of the utility matrix, depending on the algorithm.

4 The community partition matrix, which describes node membership in a disjoint manner.

By using this generalisation of unidimensional community detection algorithms, Tang et al. (2011) proposed an application to multidimensional networks by defining four possible steps of integration, through a process that begins with individual adjacency matrices for each dimension, and results in a unique community partition, combining the information of multiple types of connections.

The first integration method proposed by Tang et al. (2011) consists of treating the multidimensional network as a unidimensional network, by combining the adjacency matrices for the different types of connections into a single matrix. The authors described a single method for doing network integration, simply by calculating the average of the multiple adjacency matrices. In this work, we use the arithmetic average to calculate the integrated adjacency matrix, however, defining a network integration strategy should be a matter of creating an edge weight integration scheme – for example, in a scenario where different dimensions are used to model types of connections with dissimilar strengths, we could use a weighted average, assigning different percentual weights to the different dimensions. Similarly, the proposed utility integration method is based on a matrix average. In this case, redefining utility integration would highly depend on the type of community detection methodology used, thus a generic utility integration procedure should be calculated as an average of utility matrices. Although simple enough, network and utility integration do not always achieve the best results, and thus two other methods have been defined: structural features integration and partition integration. In fact, according to the authors, features integration has consistently achieved the best results for the tested networks. This type of integration cannot be properly computed as a matrix average, since at this level coordinates are not comparable among dimensions. Tang et al. (2011) show that the average structural features matrix, after applying a transformation $w^{(i)}$ to each dimension, is proportional to the top left singular vector of the matrix resulting of the concatenation of the structural features for each dimension. Thus, structural features integration can be

achieved simply by concatenating the structural features $S^{(i)}$ for each of the p dimensions: $X = [S^{(1)}, S^{(2)}, \cdots, S^{(p)}]$ and extracting the top left singular vectors of the resulting matrix X. The final integration methodology proposed by the authors happens at the partition level, after the adjacency matrices for each dimension have been independently used for community detection, and a partition matrix for each dimension has been obtained. Partition integration was studied as a consensus clustering problem based on hard ensemble clustering techniques. Three methodologies were proposed to compute the resulting matrix: cluster-based similarity partitioning algorithm (CSPA), hypergraph partition algorithm (HGPA), and meta-clustering algorithm (MCLA) – refer to Strehl (2003) for an overview on cluster ensemble methodologies. Given its simplicity, only the CSPA algorithm was detailed by Tang et al. (2011), even though the two other algorithms are computationally less expensive. Additionally, the authors proposed a simplistic approach they called partition feature integration, where the partitions $H^{(i)}$ for each dimension are aggregated: $Y = [H^{(1)}, H^{(2)}, \cdots, H^{(p)}]$ and the top singular vectors of the resulting matrix Y are reclustered using the k-means algorithm to compute the final partition matrix.

As proposed by Tang et al. (2011), we implemented network and utility integration as a matrix average. For simplicity, the structural features integration was done through the concatenation of the structural features matrix of each dimension, followed by the extraction of the top left singular vectors, and the application of the k-means algorithm to obtain the community partition matrix. Again, for simplicity, we implemented partition integration by using the most straightfoward approach proposed by Tang et al. (2011) – we aggregated the partition matrices and simply applied the k-means algorithm to the top left singular vectors of the resulting matrix, treating it as a feature representation of nodes.

2.2.1 Example

Let us assume a multidimensional network with two dimensions (Figure 1), defined by the following set of adjacency matrices $\mathcal{A} = \{A^{(1)}, A^{(2)}\}$. In the following steps, we exemplify how the modularity maximisation methodology (Newman's method) can be used as our community detection algorithm, and then we integrate the two dimensions at the utility matrix level, which for this particular community detection methodology corresponds to the modularity matrix $M = B$.

1 In Step 1, we present the adjacency matrices $A^{(1)}$ and $A^{(2)}$ for each dimension:

$$A^{(1)} = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad A^{(2)} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

2 In Step 2, we calculate the respective utility matrices $M^{(1)}$ and $M^{(2)}$, corresponding to the modularity matrices $B^{(1)}$ and $B^{(2)}$:

$$M^{(1)} = \begin{bmatrix} -0.333 & 0.833 & -0.167 & 0.667 & 0.000 & 0.000 \\ 0.833 & -0.083 & -0.083 & -0.197 & 0.000 & 0.000 \\ -0.167 & -0.083 & -0.083 & 0.833 & 0.000 & 0.000 \\ 0.667 & -0.167 & 0.833 & -0.333 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \end{bmatrix}$$

$$M^{(2)} = \begin{bmatrix} -0.083 & -0.167 & 0.917 & 0.000 & -0.083 & 0.083 \\ -0.167 & -0.333 & -0.167 & 0.000 & 0.833 & 0.883 \\ 0.917 & -0.167 & -0.083 & 0.000 & -0.083 & 0.083 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ -0.083 & 0.0833 & -0.083 & 0.000 & -0.083 & -0.083 \\ -0.083 & 0.833 & -0.083 & 0.000 & -0.083 & -0.083 \end{bmatrix}$$

3 In Step 3, we integrate the utility matrices by calculating their average:

$$M = \begin{bmatrix} -0.208 & 0.333 & 0.375 & 0.333 & -0.042 & -0.042 \\ 0.333 & -0.208 & -0.125 & -0.083 & 0.417 & 0.417 \\ 0.375 & -0.125 & -0.083 & 0.417 & -0.042 & -0.042 \\ 0.333 & -0.083 & 0.417 & -0.167 & 0.000 & 0.000 \\ -0.042 & 0.417 & -0.042 & 0.000 & -0.042 & -0.042 \\ -0.042 & 0.417 & -0.042 & 0.000 & -0.042 & -0.042 \end{bmatrix}$$

At this point, the problem has been reduced to the problem of community detection in a unidimensional network, starting from the utility matrix M – we emphasise that these levels were defined in the unified view that Tang et al. proposed.

4 In Step 4, we extract the structural features matrix S by aggregating the ℓ eigenvectors with the highest eigenvalues in the utility matrix. Each eigenvector represents a column of S. We use $\ell = 3$ structural features since Tang et al. (2011) showed that this parameter does not have a high impact on the results as long as a reasonably large value is chosen. Given we are working with a six node network, which can have up to six distinct eigenvectors, using three structural features corresponds to 50% of the total features available, which is a large value in this context:

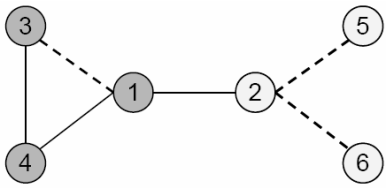
$$S = \begin{bmatrix} -0.513 & 0.256 & 0.000 \\ 0.039 & 0.689 & 0.000 \\ -0.633 & -0.040 & 0.000 \\ -0.563 & 0.017 & 0.000 \\ 0.092 & 0.478 & -0.707 \\ 0.092 & 0.478 & 0.707 \end{bmatrix}$$

5 Finally, in Step 5, we apply the *k*-means algorithm using the lines of *S* as the data points and $k = 2$ to obtain the disjoint partition matrix *H* for the two main communities in our example network:

$$H = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

The resulting communities are defined by the node sets $C_1 = \{1, 3, 4\}$ and $C_2 = \{2, 5, 6\}$, which are depicted in Figure 1 using different shades of grey.

Figure 1 Example network with two distinct dimensions (solid and dashed edges), and two identifiable communities



2.3 Ontology-based named entity recognition

We have done work concerned with the identification of language-independent contextual supernodes on co-reference networks (Devezas and Figueira, 2012). In this work, we used community detection methodologies to establish news context based on a unidimensional network where nodes represented named entities and edges represented connections between pairs of entities co-referenced in a common document. A different network, where nodes represented documents and edges represented entities referenced in a pair of documents, was also studied, but is less relevant for the scope of this work.

Next, we briefly describe the named entity recognition methodology that we used prior to the creation of the entity co-reference network. This methodology is common to the work we present here, and is used in the identification and building of the multidimensional network of entities we analyse in the following sections.

We take advantage of the Wikipedia knowledge, structured using the DBpedia ontology (Bizer et al., 2009), to preselect a set of classes and subclasses associated with the type of entities we want to identify. As each resource in DBpedia contains a label for each entity, in several languages, we can map an entity to a single URI independently of the language it was originally written in. After preselecting the classes from the DBpedia OWL file, we query the SPARQL endpoint to get lists of multilingual entities that are locally cached within a relational database. The problem can then be reduced to string matching using a finite set of patterns, in this case a list of entities. This can be implemented using the Aho-Corasick algorithm (Aho and Corasick, 1975), which is based on a finite state machine similar to a trie. The matching data structure is built once

and kept in memory, which allows for a quick matching of entity labels within the text of the document set, and results in a list of named entities for each document. The entities in this list are later on resolved to their corresponding URI, in order to make them language-independent. We use this same methodology to identify the entities for the three dimensions (places, people and dates) that we use in this work.

3 Analysing a multidimensional network

We analysed the community structure of a news clips network from the Breadcrumbs system using two different models. We treated connections as a single dimension, ignoring edge type, and we also treated connections as multiple, individual dimensions. In the multidimensional model, we compared the four integration phases proposed by Tang et al., in order to determine which approach was the most appropriate for the task of identifying coherent groups of news clips.

3.1 News clips network

We built a multidimensional network from a small collection of 121 news clips, gathered independently by five different people, across a period of 24 hours. They were instructed to use any of the following five news sources available in the Breadcrumbs system: Washington Post, Times, Telegraph, Guardian and Daily Mail, while covering five main topics: Libya, US Tax, World Debt Crisis, Italy Downgrading and Greece, as well as a sixth free topic. On average, the collected news clips were 118.6 ± 134.9 words long, ranging from 13 to 955 words.

We used the ontology-based named entity recognition process based on DBpedia (Bizer et al., 2009), previously described in Section 2.3, to identify three dimensions: places (countries, continents, islands and historic places); people (politicians, clerics, scientists, models, criminals and judges); and dates (identified using a small set of matching rules dependent on the month name). We then established connections based on the co-reference of the identified entities. This resulted in the three-dimensional network depicted in Figure 2 [we used the Fruchterman and Reingold's (1991) algorithm to set the layout of the graph]. The overall network, comprising the three dimensions, contained 94 nodes, representing news clips with references to entities (approximately 78% of the corpus), and 164 edges, representing entity co-references in pairs of clips. It has a diameter of 8 and the average geodesic distance between all pairs of nodes is 3.07.

Visually, the node size is directly proportional to the node's PageRank (Brin and Page, 1998), and the three dimensions are illustrated using different line types for the edges. Whenever two clips make a reference to a common place, a solid edge is drawn; when they co-reference a person, a dashed edge is drawn; when a common date is referenced, a dotted edge is drawn. As we can see from the figure, using either one of these dimensions alone would result in a rather disconnected network structure, comprising less information than the combination of the three. Table 1 shows the degree correlation for all combinations of subnetworks corresponding to the individual dimensions. As we can see, overall correlation is low, which confirms the lack of structural relation between the dimensions and therefore leads to conclude that each dimensions adds invaluable information to the resulting multidimensional network. In this case, for example, we can see that a set of moments (dotted edges) establish a strong

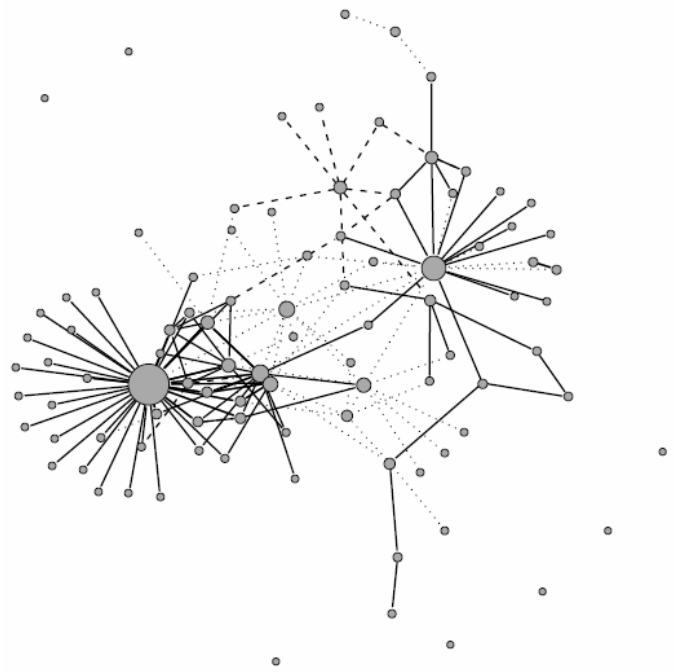
connection between two sets of places (solid edges), that would otherwise be weakly connected – were we to remove the temporal dimension, and these sets of places would have only a dashed and a solid edges creating a bridge between them. Table 2 shows the properties of the subgraphs for the individual dimensions, as well as for the multidimensional graph that combines the three dimensions. Apart from having the largest number of nodes and edges, the multidimensional graph also has the highest density and the highest clustering coefficient, which directly reflects the quality of the community structure in the network.

Table 1 Pearson correlation for the node degree of the subnetworks representing each dimension

Places/people	*Places/dates*	*People/dates*
–0.002233051	0.1198128	–0.1166953

Note: We can see a very low correlation between the degrees of the different subnetworks, which indicates that each dimension contains different information and therefore adds unique structural information to the overall multidimensional network.

Figure 2 Multidimensional network representing the co-reference of entities in news clips



Note: Solid for places, dashed for people, and dotted for dates.

3.2 Community structure

We identified and analysed the community structure of the weighted, undirected news clips network. We looked at this network both from a multidimensional and a unidimensional point of view, comparing the quality of the partitions for both cases. Our

goal here was to determine the most appropriate approach for obtaining coherent sets of clips in regard to their textual content, exclusively by using the connections between named entities in news clips. We hypothesised that by taking advantage of the additional information (i.e., the separation in multiple dimensions) contained in a multidimensional network, the resulting partition should contain a larger number of relevant communities.

Table 2 Properties of the unidimensional networks for each dimension and the multidimensional network resulting of their combination

Dimension	*Nodes*	*Edges*	*Density*	*Clustering coefficient*
Places	78	105	0.03496503	0.09396752
People	30	15	0.03448276	0.00000000
Dates	94	44	0.01006635	0.08287293
All	121	164	0.03752002	0.11070910

The multidimensional analysis of the network was based on our own Java implementation of the community detection algorithms and integration strategies proposed by Tang et al. (2011). We then used the R Project (R Development Core Team, 2011) together with the igraph package (Csárdi and Nepusz, 2006) for the unidimensional analysis, and for the study and visualisation of all of the identified partitions. For the multidimensional community detection, we used the modularity maximisation method with a fixed number of $k = 10$ communities and $\ell = k$ structural features. Fixing k facilitated the process of establishing a correspondence between communities, identified using different methods or integration phases. We used $\ell = k$ since there is a low sensitivity to the number of structural features for large values of ℓ and, since we were working with a rather small network, using $\ell = 10$ seemed appropriate. For the unidimensional community detection, we used igraph's implementation of Newman's (2006a, 2006b) leading eigenvector community detection methodology. We manually identified the number of necessary merge steps to obtain a partition of $k = 10$ communities.

4 Results

We evaluated the identified community partitions by using a twofold approach. First, we did a link-based evaluation, by measuring the quality of the partitions with the modularity score (Newman and Girvan, 2004):

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i \times k_j}{2m} \right) \delta(C_i, C_j)$$

where m is the number of edges, A the adjacency matrix, k the degree of a node, and $\delta(C_i, C_j)$ a Boolean function that returns 1 if nodes i and j belong to the same community and 0 otherwise.

For the second part of the evaluation, we collected expert human input on the individual communities, based on the semantics provided by the different groups of news clips and their textual content.

4.1 Link-based partition evaluation

We identified and compared five community partitions: four of them by using the multidimensional version of the network and the integration methods previously described, and another one by using the unidimensional version of the news clips network. Table 3 shows the modularity score of the identified community partitions. The highest modularity was achieved by treating the network as unidimensional and applying Newman's method. The multidimensional integration phases follow, with partition, network, utility and feature (from highest to lowest modularity).

Table 3 Modularity score for the partitions identified using the four multidimensional integration phases and the unidimensional method

Method	*Modularity*
Network integration	0.1940623
Utility integration	0.1801011
Feature integration	0.1180473
Partition integration	0.2093434
Unidimensional network	0.3734756

We assume that the incoherence with Tang's results, regarding the quality of the integration phases, exists partly because our network is either small or lacks a strong community structure. Furthermore, we had taken advantage of every algorithmic simplification available in the feature and partition integration phases, in order to lower the overall complexity of the problem, which might have had a negative impact on the quality of the results.

4.2 Interpreting community semantics

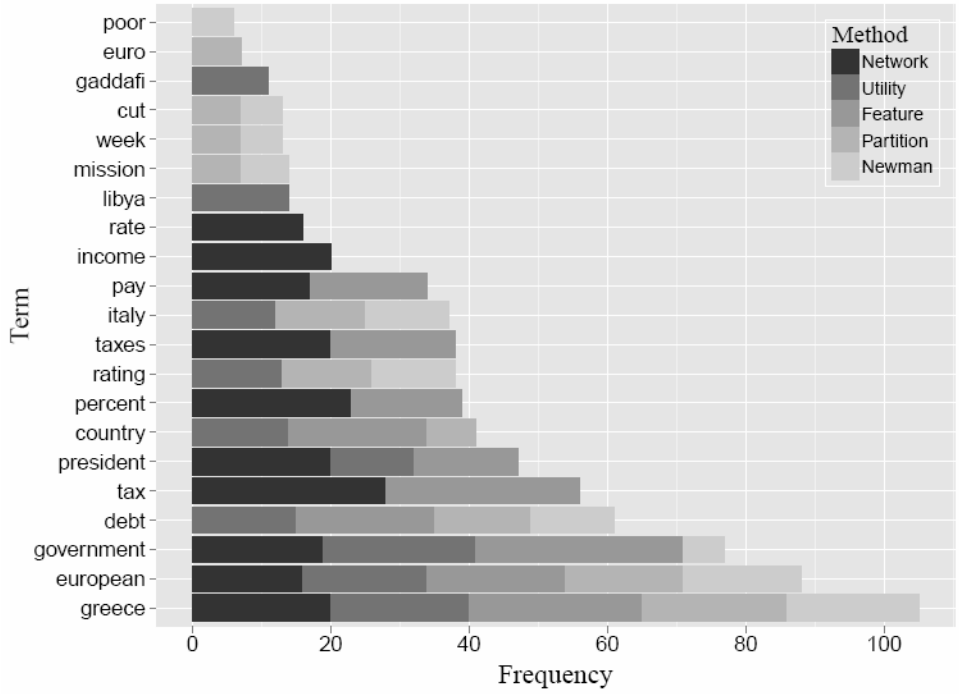
Given the previously identified inconsistencies, regarding the expected quality of the identified community structure for the integration methodologies proposed by Tang et al. (2011), we re-evaluated the quality of the different partitions now based on the semantics of each community, as imposed by the textual content of its news clips.

4.2.1 Analysing the largest community

We identified the largest community C_0 in the unidimensional version of the network, based on the partition computed by Newman's method, and used this network's community structure as our 'ground truth'. We then found the corresponding communities, in the remaining partitions, that provided the largest overlap with C_0, for each integration phase, and obtained the term frequency vectors of their aggregated news clips. Figure 3 shows the frequency of the top 10 words for each of the described communities, illustrating the differences between the methods, as well as their concordance for the most prominent topics. The overall topic seems to be the European economic crisis, however we can see the words 'libya' and 'gaddafi' in C_2 indicating a deviation from the remaining, analogous communities. Hence, even though the modularity of C_2 is higher than the modularity of C_3, a simple content-based analysis indicates that this community covers at least two distinct primary topics, slightly drifting

away from the idea that a community should represent a local group with a common main feature uniting its members. Although a higher-level topic such as politics can still be identified in C_2, when comparing community groups a greater deal of relevance should be given to a more cohesive group. Given the current case study, the community partition resulting from the utility integration phase comprises a more heterogeneous group of news clips and is therefore less suitable to achieve our goal of finding cohesive groups at a small scale.

Figure 3 Term frequency for the top 10 words in the largest community of each partition



4.2.2 Human input

We asked (I) a high school English teacher, (II) a journalist/social communication professor, and (III) a bachelor of communication to independently evaluate the quality of the communities in each partition, based on the aggregated textual content of the news clips. We first asked evaluators to assign a binary grade to each community, stating whether it is (grade 1) or is not (grade 0) a relevant group of news clips for them. This was to prepare them for a four-point scale, using values between 0 and 3, that would result in a more refined assessment. Table 4 illustrates the quality grades given by evaluators I, II and III to the 10 communities in each partition. Grades are integer numbers that vary from 0 to 3, where 0 means the evaluator was unable to identify any connection whatsoever between the news clips in a community, and 3 means the evaluator acknowledged the news clips in a community as a perfectly related and cohesive group of text.

Table 4 Human assessment of the news clips communities in each partition for evaluators I, II and III

Community	*Network*			#	*Utility*			#	*Feature*			#	*Partition*			#	*Newman*			#
	I	*II*	*III*		*I*	*II*	*III*		*I*	*II*	*III*		*I*	*II*	*III*		*I*	*II*	*III*	
1	3	3	2	3	1	0	1	51	1	0	1	57	1	0	2	17	1	1	2	14
2	0	0	0	5	2	3	3	3	3	2	3	8	2	0	1	9	0	0	2	17
3	1	0	1	13	2	1	2	3	0	0	3	1	1	0	0	14	2	1	2	7
4	2	1	3	4	3	3	3	6	2	2	3	8	3	3	3	22	1	1	2	10
5	1	0	2	52	3	3	3	2	3	3	3	2	1	0	2	15	1	1	2	9
6	3	3	3	2	3	3	3	5	1	0	1	5	2	1	1	3	3	3	3	6
7	3	1	0	3	2	1	2	4	3	1	2	3	1	2	3	3	0	0	3	1
8	3	3	3	3	2	0	2	14	2	1	1	3	1	1	2	4	0	0	2	1
9	3	3	3	7	3	0	1	3	2	0	0	5	3	0	2	3	3	2	3	20
10	3	3	3	2	3	3	3	3	1	2	2	2	2	0	2	4	2	2	3	9
Total average	22	17	23	94	24	17	23	94	18	11	19	94	17	7	18	94	13	11	24	94
		20.6667				21.3333				16.0000				14.0000				16.0000		
Eval average	133	58	177		156	64	153		135	48	132		160	79	173		143	116	224	
		122.6667				124.3333				105.0000				137.3333				161.0000		

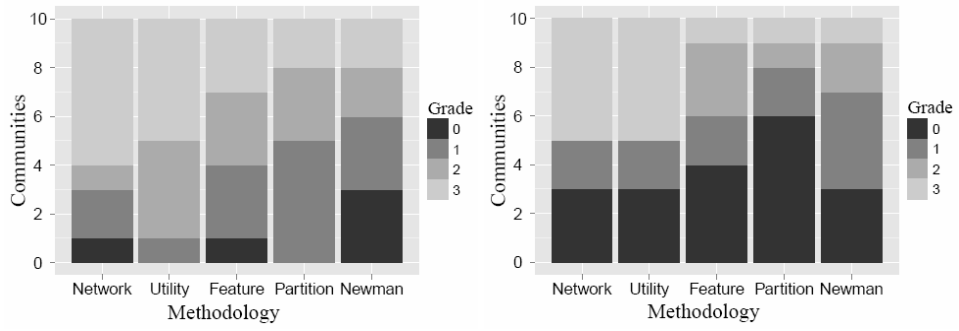
Notes: Evaluators were asked to assign a grade between 0 (worst) and 3 (best) to the individual communities, computed according to the four integration methods of the multidimensional network (*network*, *utility*, *feature*, *partition*), as well as Newman's community detection method for unidimensional networks (*Newman*). A fixed number of 10 communities was identified by each of the five methods. Here, we can see the grades assigned by evaluators I, II and III to each of the communities (rows) contained within the partitions resulting from the application of the different methodologies (columns). In the last rows, we can see the overall quality of the partitions according to two different metrics (*Total* and *Eval*).

As we can see, evaluators consistently graded communities of over 50 nodes with a low grade, meaning that news clips aggregated in those large communities did not feel cohesive to them. Evaluators also graded single-node communities using the extremities of the scale, meaning it either made no sense to the evaluator having a group of news clips composed of a single clip, or the group was immediately identified as cohesive since there were no other clips to compare the singleton to.

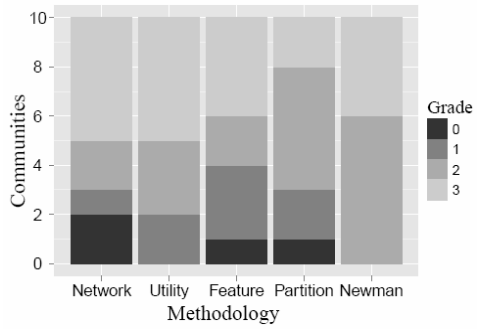
Figure 4 illustrates the fraction of communities with a grade from 0 to 3 for the identified partitions according to the five different methodologies (the four integration phases of the multidimensional community detection and Newman's leading eigenvector unidimensional community detection). Each chart depicts the assessment made by a single evaluator. The figure shows that the network and utility integration phases consistently result in the largest number of grade 3 communities and the lowest number of grade 0 communities. The remaining methodologies did not result in a consistent assessment among the three evaluators, indicating a certain ambiguity towards the aggregation of news clips as semantically cohesive communities in those partitions.

The depicted bar charts illustrate the number of communities that fall within each grade, according to the individual assessments of evaluators I, II and III, correspondingly. A bar is drawn for each methodology (*network*, *utility*, *feature*, *partition* and *Newman*) and sectioned according the frequency of each individual grade.

Figure 4 Grade distribution for the communities in each partition assigned by (a) evaluators I (b) II and (c) III



(a) (b)



(c)

5 Discussion

In this section, we analyse the results of our experiments by comparing the link-based evaluation with the content-based human evaluation, while commenting on the conformity between evaluator's assessments. We also examine the influence of community size in the measurement of the partition quality.

5.1 The quality of partitions

We compare the results of the two evaluation methods: the link-based approach, and the content-based semantic approach. The first method purely took advantage of the links between clips, using the modularity metric to measure the quality of the partitions. The second method took advantage of the textual content present in each node, to establish a semantic evaluation relying on human input. In Table 4, we have the total grade of each partition, calculated using the sum of the grades for the individual communities, and the average for the three evaluators, which was used as an indicator of quality. We also have an 'Eval' metric that illustrates the quality of a partition based on the grades assigned by the evaluator, as well as the number of nodes that belong to the communities – this was obtained through the sum of the products between grade and number of nodes (#). Again, we calculated the average for the three evaluators, which was used as a quality metric that takes higher values for large communities with a high grade. Human evaluators were not always consistent among each other. For instance, the three evaluators graded community 7 of the network integration phase with 3, 1 and 0, respectively. This community was composed of three news clips, all of which about the world debt crisis. Two of these clips were specifically about the European economic crisis, mentioning Greece, Ireland, Portugal and 2011, while the third clip was about the US Tax, being connected to the other two clips only by the year 2011. A similar situation happens for community 8 of the utility integration phase, where a clip about video games was clustered with news clips with topics such as the recapitalisation of banks or the intervention of the International Monetary Fund, solely based on the co-reference of 2008, another weak date connection. This is an indicative that using a yearly resolution for the date dimension might decrease the quality of the community structure by forming strong relationships that should otherwise be considered weak or simply removed. Dimensions selection is a fundamental step in modelling a multidimensional network for community detection and it is definitely worth investigating in the future. In Table 5, we can find the five community detection methodologies ordered according to each of the quality metrics. Based on the modularity metric (calculated in Table 3), the best partition results from the application of Newman's method to the unidimensional version of the network. However, human evaluators have consistently identified the Utility integration phase of the modularity maximisation method, when applied to the multidimensional network, as the best partition of the network (highest average grade). Given we are in our current project developing a social platform, human input is of a higher relevance to the evaluation of the system.

Community structure is a property of complex networks that is more prominent in large real-world networks. We have, however, been able to successfully identify coherent communities in a small network of 94 nodes, by taking advantage of link diversity and analysing a multidimensional version of the network. We conjecture that, even though the

two types of evaluation methods are pointing towards different outcomes, as the network grows there should be a more evident similarity of quality among the five detection methods.

Table 5 Methods ordered by the quality of their resulting community structure, for three different metrics

Rank	*Modularity*	*Human (total)*	*Human (Eval)*
1	Newman	Utility	Newman
2	Partition	Network	Partition
3	Network	Feature	Utility
4	Utility	Newman	Network
5	Feature	Partition	Feature

5.2 The influence of community size

For every identified community partition in our network, there is an exceptionally large community, either around 50 nodes or 20 nodes. This is highly dependent on the detection methodology and the network itself; the integration of dimensions should have little or no influence on the generation of such a large community. This special case raises some questions as to whether or not a high grade large community should have a higher weight on the final quality score.

We tested this by calculating the previously described 'Eval' metric. When ordering the methodologies by quality of partition, the results are similar both for the 'modularity' metric and the 'Eval' metric. However, we defend that evaluators should have had a harder time identifying a cohesive group of news clips for a larger community, than they would have had for a smaller community. This means that community size is already included in the evaluation process and did not need therefore to be combined once again with the final score. Nevertheless, this information is useful, as it becomes clear based on the 'Eval' metric that the Newman method is better when we need to ensure that the largest communities consist of higher quality news clips aggregations. We also notice that partition integration introduces improvements over the remaining integration phases, regarding the quality of the largest communities.

6 Conclusions

We studied the community structure of a weighted, undirected news clips network, where we used multiple features to establish connections between clips. We identified and compared the partitions that resulted from a multidimensional approach, where edges of different types contributed independently to the discovery of the community structure, with the partition that resulted from a unidimensional approach, where no distinction was made between edges of different types. It became clear that the use of several dimensions can contribute to the improvement of the community detection process in small networks. In order to obtain semantically coherent communities, dimensions should however be chosen carefully to reflect the characteristics responsible for a good group formation.

We found that, from a user's point of view, better groups of news clips can be identified by using the modularity maximisation methodology with the integration at the

utility matrix level, closely followed by the integration at the network (adjacency) matrix level. We also verified that, for our network of news clips, the Newman method, when applied to a unidimensional version of the network, has rather different results than the same modularity maximisation method applied to the multidimensional network when combined with the network integration phase.

The successful identification of the community structure of multidimensional networks is a fundamental step in the analysis of social networks, where the integration of multiple signals or dimensions, describing different aspects of social interaction, is becoming rather frequent.

7 Future work

As future work, we would like to replicate this experiment at a larger scale, after the dataset of news clips has been increased. Additionally, we would like to extend this procedure to encompass overlapping community detection methodologies, as well as multimodal and dynamic networks, with the goal of understanding how partitions can be improved by further introducing information to create a network model that more accurately reflects reality. We would also like to study how the quantity of features or dimensions influences the formation of community structure for our kind of network, trying to understand whether or not there is a certain limit for the number of features, where the quality of the partitions actually starts decreasing. Finally, we point out that community detection in multidimensional networks has only recently been explored, making use of classic community detection methodologies that are quickly being replaced by novel and more scalable algorithms, where the unified view and the integration strategies proposed by Tang et al. (2011) do not directly apply. Enabling the application of state of the art community detection methodologies to multidimensional networks is an open problem where new integration solutions are still waiting to be developed. Thus, this is a very promising area of research within network science and complex systems.

Acknowledgements

This work is financed by the ERDF – European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT – Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project UTA-Est/MAI/0007/2009.

References

Aho, A.V. and Corasick, M.J. (1975) 'Efficient string matching: an aid to bibliographic search', *Communications of the ACM*, June, Vol. 18, No. 6, pp.333–340.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R. and Hellmann, S. (2009) 'DBpedia – a crystallization point for the web of data', *Web Semantics: Science, Services and Agents on the World Wide Web*, September, Vol. 7, No. 3, pp.154–165.

Borg, I. and Groenen, P.J.F. (2003) 'Modern multidimensional scaling: theory and applications', *Journal of Educational Measurement*, September, Vol. 40, No. 3, pp.277–280.

Brin, S. and Page, L. (1998) 'The anatomy of a large-scale hypertextual web search engine', *Computer Networks and ISDN Systems*, Vol. 30, Nos. 1–7, pp.107–117.

Csárdi, G. and Nepusz, T. (2006) 'The igraph software package for complex network research', *InterJournal Complex Systems*, Vol. 1695, No. 1695, pp.1–9.

Devezas, J. and Figueira, A. (2012) 'Finding language-independent contextual supernodes on coreference networks', *IAENG International Journal of Computer Science*, Vol. 39, No. 2, pp.200–207.

Devezas, J., Alves, H. and Figueira, A. (2012) 'Creating news context from a folksonomy of web clipping', in *Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists 2012 (IMECS 2012)*, Hong Kong, pp.446–451.

Doreian, P., Batagelj, V. and Ferligoj, A. (2004a) 'Generalized block-modeling of two-mode network data', *Social Networks*, Vol. 26, No. 1, pp.29–53.

Doreian, P., Batagelj, V. and Ferligoj, A. (2004b) *Generalized Blockmodeling*, Cambridge University Press, Cambridge.

Edmonds, J., Raschid, L., Sayyadi, H. and Wu, S. (2010) 'Exploiting social media to provide humanitarian users with event search and recommendations', in *Proceedings of the 7th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2010)*, Seattle, USA, pp.1–5.

Figueira, A., Ribeiro, P., Leal, J.P., Zamith, F., Cunha, E., Francisco-Revilla, L., Ribeiro, H., Silva, A., Pinto, M., Alves, H., Devezas, J., Santos, M. and Cravino, N. (2009) 'Breadcrumbs: a social network based on the relations established by collections of fragments taken from online news' [online] http://breadcrumbs.up.pt (accessed 19 January 2012).

Fortunato, S. (2010) 'Community detection in graphs', *Physics Reports*, Vol. 486, Nos. 3–5, pp.75–174.

Fruchterman, T.M.J. and Reingold, E.M. (1991) 'Graph drawing by force-directed placement', *Software: Practice and Wxperience*, Vol. 21, No. 11, pp.1129–1164.

Gargi, U., Lu, W., Mirrokni, V. and Yoon, S. (2011) 'Large-scale community detection on youtube for topic discovery and exploration', in *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011)*, pp.486–489.

Jin, D., Yang, B., Baquero, C., Liu, D., He, D. and Liu, J. (2011) 'A Markov random walk under constraint for discovering overlapping communities in complex networks', *Journal of Statistical Mechanics: Theory and Experiment*, May, Vol. 2011, No. 5 p.05031.

Kushima, M., Araki, K., Suzuki, M., Araki, S. and Nikama, T. (2012) 'Text data mining of the electronic medical record of the chronic hepatitis patient', in *Proceedings of the International MultiConference of Engineers and Computer Scientists 2012 (IMECS 2012)*, Vol. 1, Kowloon, Hong Kong.

Leskovec, J., Lang, K.J. and Mahoney, M. (2010) 'Empirical comparison of algorithms for network community detection', in *Proceedings of the 19th International Conference on World Wide Web*, ACM, pp.631–640.

Leskovec, J., Lang, K.J., Dasgupta, A. and Mahoney, M.W. (2008) 'Statistical properties of community structure in large social and information networks', in *Proceeding of the 17th International Conference on World Wide Web*, ACM, pp.695–704.

Leung, I.X.Y., Hui, P., Lio', P. and Crowcroft, J. (2008) 'Towards real-time community detection in large networks', *Physical Review E*, August, Vol. 79, No. 6, p.10.

Mucha, P.J., Richardson, T., Macon, K., Porter, M.A. and Onnela, J-P. (2010) 'Community structure in time-dependent, multiscale, and multiplex networks', *Science*, November, Vol. 328, No. 5980, pp.876–878.

Neubauer, N. and Obermayer, K. (2009) 'Towards community detection in k-partite k-uniform hypergraphs', in *Proceedings of the NIPS 2009 Workshop on Analyzing Networks and Learning with Graphs*, pp.1–9.

Newman, M.E.J. (2006a) 'Finding community structure in networks using the eigenvectors of matrices', *Physical Review E*, Vol. 74, No. 3, p.36104.

Newman, M.E.J. (2006b) 'Modularity and community structure in networks', *Proceedings of the National Academy of Sciences*, June, Vol. 103, No. 23, p.8577.

Newman, M.E.J. and Girvan, M. (2004) 'Finding and evaluating community structure in networks', *Physical Review E*, Vol. 69, No. 2, p.026113.

Ohsawa, Y. and Benson, N.E. (1998) 'KeyGraph: automatic indexing by co-occurrence graph based on building construction metaphor', in *Proceedings of the IEEE International Forum on Research and Technology Advances in Digital Libraries*, pp.12–18.

Palla, G., Derenyi, I., Farkas, I., Vicsek, T. and Derényi, I. (2005) 'Uncovering the overlapping community structure of complex networks in nature and society', *Nature*, Vol. 435, No. 7043, pp.814–818.

R Development Core Team (2011) 'R: a language and environment for statistical computing', in *R Foundation for Statistical Computing*, Vienna, Austria.

Sayyadi, H., Hurst, M., Maykov, A. and Microsoft Livelabs (2009) 'Event detection and tracking in social streams', in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007)*.

Strehl, A. (2003) 'Cluster ensembles – a knowledge reuse framework for combining multiple partition', *The Journal of Machine Learning Research*, 1 March, Vol. 3, pp.583–617.

Šubelj, L., Bajec, M. and Subelj, L. (2011) 'Generalized network community detection', Arxiv preprint arXiv:1110.2711.

Tang, L., Wang, X. and Liu, H. (2011) 'Community detection via heterogeneous interaction analysis', *Data Mining and Knowledge Discovery*, August.

von Luxburg, U. (2007) 'A tutorial on spectral clustering', *Statistics and Computing*, Vol. 17, No. 4, pp.395–416.

Xie, J., Kelley, S. and Szymanski, B.K. (2011) 'Overlapping community detection in networks: the state of the art and comparative study', Arxiv preprint arXiv: 1110.5813, October, Vol. 5, pp.1–30.