A Structural Analysis of topic ontologies
E Xamena, NB Brignole, AG Maguitman
Information Sciences 421, 15-29

A Structural Analysis of Topic Ontologies

Eduardo Xamena[a,b,c], Nélida Beatriz Brignole[b,d,e], Ana Maguitman[c,d]

[a] *Facultad de Ciencias Exactas - UNSa - Universidad Nacional de Salta - Av. Bolivia 5150, Salta, Argentina.*
[b] *LIDeCC - Laboratorio de Investigación y Desarrollo en Computación Científica.*
[c] *Knowledge Management and Information Retrieval Research Group, ICIC CONICET-UNS.*
[d] *DCIC-UNS - Departamento de Ciencias e Ingeniera de la Computacin - Universidad Nacional del Sur - San Andrés 800, Bahía Blanca, Argentina.*
[e] *Planta Piloto de Ingeniería Química, UNS-CONICET - Cno la Carrindanga km 7, Bahía Blanca, Argentina.*

Abstract

DMOZ is the largest human-edited topic ontology available on the Web. This article studies the structural properties of the DMOZ graph. A number of global and local properties of this graph and the subgraphs resulting from isolating edges of different types are examined by means of metrics commonly used in complex network analysis. In particular, we investigate the presence of various features that characterize small-world networks. This analysis is complemented by examining other characteristics of the graphs such as connectivity and centrality measures. The connectivity and centrality patterns are further studied by means of visualizations of the graphs' k-core decomposition and a selection of strongly connected components. Several non-trivial regularities that are also encountered in other artificial and natural complex networks provide a general picture of this large human-edited topic ontology. This analysis is of major pragmatic interest as it allows a better understanding of notions such as navigability among topics, hierarchical structure and topic cohesiveness, which are of great importance in the design of topic ontologies.

Keywords: topic ontologies, topological analysis, complex networks

Corresponding author email: examena@plapiqui.edu.ar

1. Introduction

Ontologies are structures commonly used to capture knowledge about certain areas by providing relevant topics or concepts and relations between them. A general topic ontology such as DMOZ (Directory Mozilla), historically known as ODP (Open Directory Project), is a complex structure that reflects the collective knowledge of the ontology editors about a broad range of topics. Revealing information about structural aspects of this ontology can provide useful insights on the nature of topic connectivity, topic importance, topic relevance and topic similarity, among other useful concepts, conferring a unique opportunity to address the important cognitive problem of understanding "the landscape of topics" as realized by a large number of human editors.

The DMOZ is a collaborative classification of websites. The topical structure made up from this classification can be represented as a big graph or ontology. In the DMOZ topic ontology, topics are represented as nodes in a tree-structured hierarchy, with "is-a" connections determined by topic-inclusion relations. In addition, DMOZ admits cross-links representing "symbolic" connections to allow for topics with multiple parents. Finally, another type of relation represented by "related" links allows to connect related topics that are not involved in a class-inclusion relation. While the tree-structured hierarchy imposes strong constraints on the general organization of the DMOZ ontology, the "symbolic" and "related" connections loose up these constraints and offer the possibility of integrating the taxonomical component of DMOZ with more general components, resulting in less restricted connectivity patterns when analyzed as a whole.

Network analysis constitutes a powerful tool for inferring several properties on datasets arising from a wide range of areas. From the topological structure of the web [9] to the analysis of the economy of a country [16], network properties reveal many important features of the represented models. These properties have important implications on the robustness, navigability, and cohesiveness of the networks. Large volumes of linked data can be analyzed from a Complex

Network perspective, with application to information retrieval. A structural study of a big corpus, made up of interconnected documents or other kinds of information entities, could help on finding useful information about the semantic relations existing among these entities. Graph representations have proved to be an effective and efficient way for structural semantic similarity calculations [28]. The structure of semantic networks constructed from word associations has been widely studied in cognitive science [39, 7, 31], with application in several areas such as the assistance of people with the anomic aphasia disease [34].

The study presented here focuses on analyzing the network topology of the DMOZ graph in its pure form. It also analyses the network topology of the subgraphs of DMOZ corresponding to edges of the three different types involved in this ontology, namely taxonomical, symbolic and related edges. The analysis is carried out by computing various complex network metrics, such as node degree, local clustering coefficient, average shortest path length, and diameter of the network, allowing to draw interesting conclusions about non-trivial regularities present in the analyzed graphs. To the best of the authors' knowledge this article provides the first large-scale analysis of a topic ontology graph from a complex network perspective.

2. Background

In this section some graph-theoretic concepts are briefly described, in particular those that relate to the analysis carried out on the DMOZ structure. Then, we describe various measures and tools that have been adopted to complete the analysis reported in this article.

2.1. DMOZ as a graph

The DMOZ project is a large directory of websites organized by topics. The main component of this directory is a hierarchical structure, the DMOZ taxonomy. Websites are added to the directory by assigning them to existing topics from the taxonomy. Besides its hierarchical structure, DMOZ contains

other kinds of links between topics, as is the case of "symbolic" and "related" links. "Symbolic" links correspond to alternative classifications that escape from the taxonomy, and have to be included in the directory. "Related" links are used to connect topics that are associated according to some criterion whenever such relation is not expressed as a taxonomic or symbolic relation. More formally, the structure of DMOZ can be represented as a directed graph $G = (N; E)$ with a set of nodes N and a set of edges E. Each node in N represents a topic containing documents, and every edge of E connects two nodes of N. The set of edges E is made up of three classes of links between topics:

- class T, corresponding to the hierarchical component of the ontology,
- class S, reflecting the non-hierarchical "symbolic" cross links, and
- class R, representing the "related" cross links, also organized in a non-hierarchical fashion.

These three types of links give rise to the *T-subgraph*, *S-subgraph* and *R-subgraph*, respectively. Each of these subgraphs will be analyzed as independent networks as well as jointly.

Figure 1 illustrates a portion of the structure of the DMOZ ontology graph, showing the three types of links.

2.2. Structural Analysis of Graphs

This section reviews some concepts, measures and algorithms that we have adopted to analyze the most salient properties of the DMOZ ontology graph.

2.2.1. Connectivity and centrality measures

Several connectivity and centrality measures commonly used for complex network analysis can be applied to the DMOZ graph, offering a means to assess topic importance and relevance among topics. Next, we describe the connectivity and centrality measures used in this work.

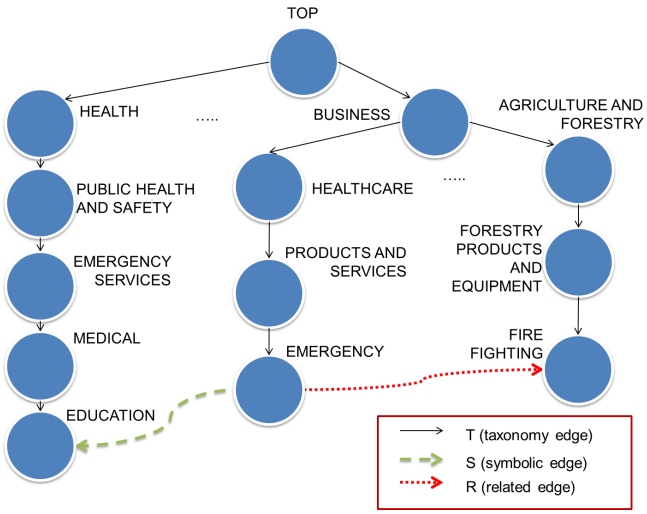

Figure 1: Portion of the DMOZ ontology.

- **Graph density**: The density of a graph is the proportion of edges actually present in a graph with respect to the number of possible links that could be established between the nodes of the graph. This measure is computed on a graph G = (N,E) as follows [14]:

$$Density(G) = \frac{|E|}{|N|(|N|-1)}.$$

- **Diameter**: The diameter of a graph is characterized by the largest distance between any two nodes, where the distance between a pair of nodes is the length of the shortest path between them [22].

- **Characteristic Path Length (CPL)**: Also known as *Average path length*, is the mean length of all the shortest path lengths in the graph. This mean value is very representative for several features of a graph, as is the case of the *"Small World property"* [41], which is described in the next subsection. The formula for computing the CPL l of a graph G=(N, E) is:

$$l(G) = \frac{1}{|N| * (|N|-1)} \sum_{i \in N} \sum_{j \in N \setminus \{i\}} spl(i,j),$$

where $spl(i,j)$ is the shortest path length for nodes i and j.

- **Connectivity length (CL)**: In [29] an alternative to the CPL is proposed. Instead of using the arithmetic mean of the shortest path lengths, the harmonic mean is used. This measure attempts to address the problem of disconnected nodes. If a node is not reachable from another one, then the distance between them is ∞. Supposing that $\infty^{-1} = 0$, the connectivity length for a graph G = (N,E) is computed as:

$$D(G) = \frac{|N|(|N|-1)}{\sum_{i,j \in N} \frac{1}{d(i,j)}}.$$

- **Local Clustering Coefficient**: This measure is defined as a degree of interconnection between the neighbors of a node [41]. For the calculation of this coefficient in directed graphs, the number of real edges between neighbors of the corresponding node i is divided by the total amount of possible edges between them, according to the next formula:

$$C_i = \frac{|\{e_{jk} : j,k \in N_i; e_{jk} \in E\}|}{|N_i|(|N_i|-1)},$$

where N_i is the set of neighbors of node i, and e_{jk} is an edge between nodes j and k. This coefficient lies on the interval [0, 1], and reflects the proportion of edges between neighbors present in the graph induced by the neighbors of node i.

- **Betweenness Centrality**: If a node i plays an important role in the graph structure, it is likely to be situated in a central place of the network. In order to measure this property, the *betweenness centrality* (BC) degree is the proportion of shortest paths between every pair of nodes in the graph that pass through node i. This measure was first proposed in [19], and an efficient algorithm for its computation has been developed in [8]. The betweenness centrality of a node i is computed as follows:

$$bc_i = \sum_{j \neq i \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}},$$

where σ_{jk} is the total number of shortest paths from a node j to another node k, and $\sigma_{jk}(i)$ is the number of shortest paths from node j to node k that pass through i.

- **Closeness Centrality**: The closeness of a node i is defined as the reciprocal of the sum of distances to that node from every other node in the network [5]. When a node is relatively central in a network, this measure is expected to be higher. As for betweenness centrality, this measure is defined over the shortest paths between every pair of nodes. The existence of nodes that cannot reach i represents a problem for computing closeness centrality, but it is overcome by considering as zero those distances [6]. The formula is given by:

$$cl_i = \frac{1}{\sum_{d(j,i)<\infty} d(j,i)},$$

where $d(j,i)$ is the shortest path distance between nodes j and i.

- **Harmonic Centrality**: Another way for evaluating centrality of a node i is to compute the harmonic mean over all distances from every node j to node i, for the set of co-reachable nodes of i [29]. Then, this measure is calculated as follows:

$$hc_i = \sum_{d(j,i)<\infty, j\neq i} \frac{1}{d(j,i)}.$$

According to this formula, the longer are the distances to node i, the smaller is the centrality value.

- **Lin's index for Closeness Centrality**: In order to obtain a more appropriate definition of closeness centrality, an alternative index to measure this value was presented in [27]. This alternative measure weights closeness by means of the square of the number of co-reachable nodes. Thus, the definition for this measure is:

$$lin_i = \frac{|\{j : d(j,i) < \infty\}|^2}{\sum_{d(j,i)<\infty, j\neq i} d(j,i)}.$$

Instead of considering the inverse of a sum of distances, this notion of centrality is based on the inverse of the average distance. This way, a normalization in the value of closeness is attained. The square in the

formula allows to weight more strongly those nodes with a larger set of co-reachable nodes.

- **Degree distribution**: As known in graph theory, the degree k of a node i in an undirected graph reflects the number of edges or connections the node i has to other nodes and is computed as follows [20]: $k_i = \sum_j a_{ij}$, for all the nodes j connected to node i. In the case of directed graphs, this degree is indicated with two numbers, namely the in-degree k^{in} and the out-degree k^{out}: $k_i^{in} = \sum_j a_{ji}$, and $k_i^{out} = \sum_j a_{ij}$, for all the nodes j connected to node i. The degree distribution *P(m)* for undirected graphs is defined as the probability that a node is linked to m nodes. For the case of directed graphs, the in-degree distribution $P_{in}(m)$ and the out-degree distribution $P_{out}(m)$ are defined as the probabilities for any node of having m incoming or outgoing links respectively. Given a directed graph G =(N,E), these distributions are computed as follows:

$$P_{in}(m) = \frac{\text{Number of nodes with in-degree } m}{|N|},$$

and

$$P_{out}(m) = \frac{\text{Number of nodes with out-degree } m}{|N|}.$$

2.2.2. The "Small World" property

The *Small-World* property is an indicator of the topology of a network that provides information about its robustness, propagation speed, computational power and synchronizability [41]. These networks have a low CPL *l(G)*, turning any node reachable in relatively few steps from any other node. Another representative feature of Small-World networks is a high level of *clustering coefficient*, yielding high connectivity for the entire graph. The networks that exhibit this feature could be placed on an intermediate point between regular lattices and random graphs, as empirically demonstrated in [41].

An important issue on the small-world analyses from Watts and Strogatz is the requirement of *total connectedness*. If there are isolated groups of nodes,

measures like the *CPL* could not represent the complete network behavior correctly. The work of Marchiori [29] introduces the use of harmonic means instead of classical geometric means. With this change on the measures, non-connected networks might be studied properly. In the present article, we perform the small-world analysis on DMOZ Ontology employing both the Watts-Strogatz and Marchiori-Latora methods.

2.2.3. Strongly connected components

The structure of complex networks often involves sets of nodes interconnected by paths, as well as pairs of nodes that are unreachable from each other. In a broader view, some groups of connected nodes could be found. Such groups are known as connected components in graph theory. Particularly for directed graphs, a strongly connected component (SCC) consists of a group of nodes with the following property:

> For every pair of nodes (i, j) that belong to the same strongly connected component, there is a path from i to j.

The algorithm proposed by Tarjan [40] performs the detection of SCC's by means of the identification of cycles in the graph. For the particular case of DMOZ, the existence of links between nodes of different branches of the taxonomy, i.e. symbolic or related links, can give rise to SCC's. Figure 2 shows an example of a small taxonomy with cross edges. In this example, two SCC's are highlighted.

2.2.4. Power-law distributions

Power-law distributions can be modeled with the following function of probability distribution [33]:

$$p(x) = Cx^{-\alpha}.$$

In this paper the negative exponent α denotes the *scaling exponent* of each analyzed power-law distribution. Particularly, when this value is greater than 2, the distribution is said to have a well-defined mean, and if it is higher than 3,

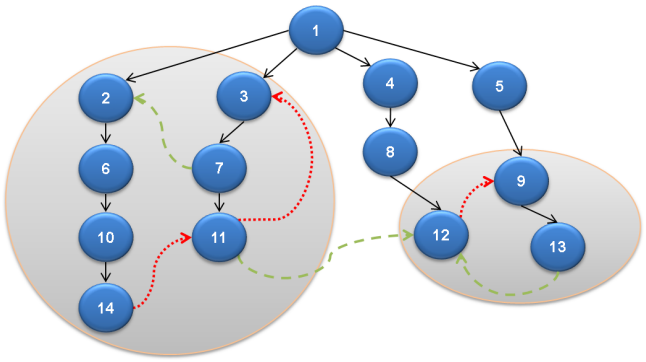

Figure 2: An example of the detection of SCC's in a DMOZ-like graph

the associated phenomenon also has a finite variance. When the node degrees of a network follow a power-law distribution, such network is said to be *Scale-free*.

2.2.5. Visual Analysis

The visual representation of the DMOZ ontology allows to identify a number of interesting features that otherwise would remain unnoticed. A useful tool for the visualization of large graphs is *Large Networks Visualization* (LaNet-Vi, [2]). This tool generates visualizations of undirected graphs, employing a *k-core decomposition algorithm*. The k-core decomposition [38] systematically identifies layers of nodes with equal degree. The most outer layer of a k-core decomposition will be made up by the least connected nodes, and the central layer by the most connected ones. It is possible to define an alternative degree value that represents the degree of a node after the layers external to that node are identified. This degree, referred to as *shell-degree*, takes a value smaller than or equal to the node's original degree.

In Figure 3 the topology of a network is graphically represented by means of the LaNet-Vi tool. The shell-degrees resulting from the k-core decomposition are shown in the legend at the right-hand side of the figures. The nodes are displayed over different circles, according to several criteria. For instance, nodes included in the same circle belong to the same shell and the size of a node determines its degree.

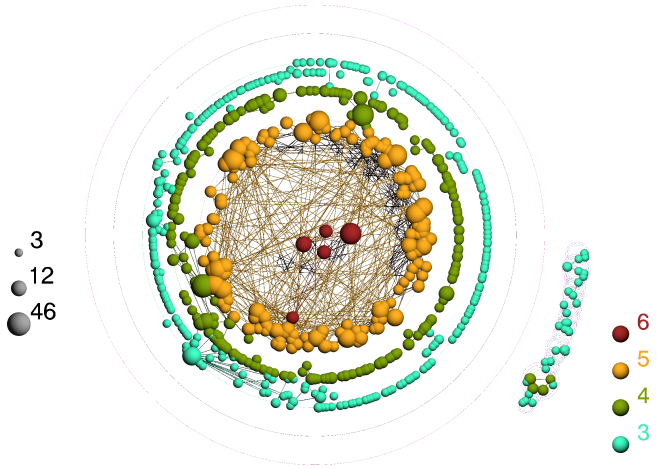

Figure 3: Graphical representation of a k-core decomposition

Gephi [4] is another software tool that enables to compute several metrics on graphs and to apply different graph layout techniques for visualizing nodes and edges. In this work, Gephi is employed to render visualizations of some communities of nodes of the DMOZ graph, according to the identified SCC's.

3. Related work

The analysis of the global structure of different datasets represented as networks provides a wide-range perspective of their static and dynamic aspects. This section presents a literature overview of approaches aimed at analyzing different kinds of networks from a complex network perspective, particularly focusing on semantic networks and ontologies.

An interesting research field in psychology and neuroscience that has been studied from the complex networks perspective is the way human and animal brains work. Different studies have recognized that brain networks share certain key organizational principles with other complex networks, such as short path length, high clustering coefficient, hierarchical structure, and power-law degree distribution ([26, 18]). Some works study the trade-offs from multiple constraints involved in the organization of neural systems, such as Bullmore

[11]. The work of Bullmore and Sporns [10], discusses how the organization of complex brain networks is conserved over different scales and for different species. Many other systems and processes have been studied as complex networks. These include the World Wide Web [1], the Internet [36], peer-to-peer networks [25], science collaboration networks [32], metabolic networks [24], epidemic processes [35], and the economic structure of countries [16], among others. An extensive review of real-world phenomena analyzed as complex networks is presented in [15].

The nearest research area for this paper is comprised by semantic networks and ontologies. Understanding the way knowledge emerges, evolves and is retrieved by the human brain has been a long-standing research focus in cognitive science. Insights into these problems could constitute a milestone in the development of new methods and techniques for more effective information retrieval and human-computer interaction, among a wide range of purposes. A useful starting point for understanding knowledge organization phenomena in human subjects has been to study semantic word association networks, at aggregative and individual levels, from a complex networks perspective. The work of Steyvers and Tenenbaum [39] tries to explain the process of semantic growth in terms of new generation models of specific types of networks. The proposed models offer an explanation for the large-scale structural features of the semantic networks under analysis, such as their small-world nature and degree distributions, among others. However, the authors argue that the process of network growth that governs the evolution of word association networks is different from the classical models proposed by other authors, such as the small-world network generation process described by Watts and Strogatz [41] or the preferential attachment model of Barabási and Albert [3]. Nevertheless, while Steyvers and Tenenbaum [39] state that the semantic networks studied are scale-free, Morais et al. [31] have argued that node-degree distributions different from the general power-law distribution fit better in some cases. Particularly, statistical tests have determined that the power-law with exponential cutoff distribution [13] provides a better fit for some networks. Morais et al. [31] show results for individual se-

mantic networks that are in accordance with these findings, and state that the growing network models proposed are in good agreement with the structure of the semantic memories generated by the evaluated individuals. But there is still a long way to go in semantic networks and brain connectivity research, in order to uncover the issues of human cognitive processes. Given that DMOZ is a human-conducted classification of topics, it could serve as another aggregative semantic network that reflects collective knowledge on topical structures, shedding light to the cognitive processes studied.

Besides the spread of semantic network studies arising from aggregate word association, several works have applied the complex network point of view to study domain-specific ontologies. Particularly, "folksonomies" have captured the attention of researchers, as they are intended to generate better semantic networks by involving the social dimension in specific contexts. In [12] the small-world, scale-free and several network properties are uncovered for specific folksonomies. Mika [30] compares the accuracy of folksonomies against text-mining-generated ontologies, recognizing the most central nodes and the most cohesive clusters in terms of betweenness centrality and clustering coefficient, and Hoser et al. [23] perform studies over ontologies, analyzing centrality and connectedness of nodes for identifying concepts that could be unified. Besides considering associations of concepts, Gueret et al. [21] propose a framework to analyze and enhance the quality of the links within a document network by means of topological and similarity measures, increasing clustering and decreasing the characteristic path length.

4. Analysis

The measures described in the Background section were applied to the DMOZ ontology structure. According to the results, several interpretations are provided regarding the topological properties of a topical ontology.

4.1. Data set

The analyzed data set comprises the basic DMOZ ontology, with the three types of links detailed in the Background section. Many efficient algorithms have been implemented for processing the large structures associated with the DMOZ graph. The employed ontology comprises a set of 571,148 topics. Table 1 summarizes the number of edges for the three types of links in the DMOZ ontology.

Table 1: Number of edges of each type in the DMOZ Ontology

Type of link	Number of edges
T (Main Taxonomy)	571,147
S (Symbolic)	545,805
R (Related)	380,264

4.2. Strongly Connected Components

The SCC's of the DMOZ graph and the S and R components were computed using the algorithm of Tarjan [40] explained in the Background section. The taxonomy T was not considered as an independent subgraph in this study given that, as it is a tree structure, it does not present cycles and hence no SCC's could be found. The goal of this analysis is to determine whether the distribution of the sizes of the SCC's follows a power law, and to verify the existence of a big connected component, among other properties.

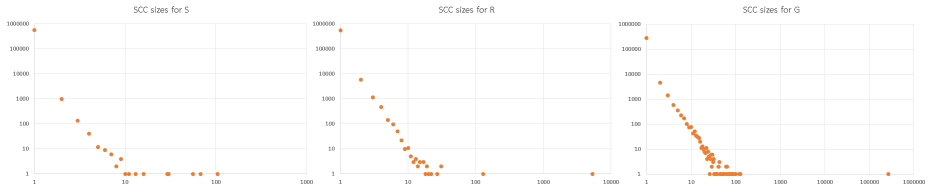

Figure 4: Distribution of SCC sizes for DMOZ graphs in log-log scale

Figure 4 shows the size distribution of SCC's in log-log scale for the *S*- and *R-subgraphs* and for *G*. The scaling exponents of the approximation formulae for the SCC distributions are reported in table 2. This table also shows the size of the greatest SCC for each graph, and the number of one-node components. It is interesting to note that the largest SCC component of *G* contains nearly half of the DMOZ topics. Different from *G*, the *R*- and *S-subgraphs* do not exhibit the presence of a considerably big SCC. The scaling exponent is larger for *R* than for *G*, but this could be a consequence of the absence of a larger SCC in *R*. The highest number of isolated nodes corresponds to the component *S*, which also has the lowest size for the largest SCC. The *S-subgraph* is the only graph that presents a considerably high scaling exponent for the distribution of SCC sizes. *R* and *G* have scaling exponents lower than 2, and this indicates that their SCC distributions do not have a well-defined mean. However, the number of nodes that are isolated or belong to the largest SCC in each of these two graphs causes a distortion in the real nature of their distributions. If the largest SCC and the isolated nodes are not considered in the graphs *R* and *G*, the scaling exponents become 2.502 and 2.135, respectively, and it is observed that the data fit power-law distributions with better defined parameters.

Table 2: Properties of SCC's for DMOZ graphs

Graph	Scaling exponent	Size of biggest SCC	Number of isolated nodes
S	2.202	104	568,080
R	1.410	5,314	545,948
G	1.100	267,541	274,552

4.3. Connectivity and Centrality measures

Global measures —density, diameter, average shortest path length and connectivity length— and local measures —local clustering coefficient, betweenness

centrality, closeness centrality, harmonic centrality and Lin's index— were computed for the graphs and are reported and analyzed in the next subsections.

4.3.1. Graph Density

The second column of Table 3 shows the densities of the DMOZ graphs. Densities are low for the four analyzed graphs, as only a small number of links exist compared to the maximum possible number of links between topics in DMOZ.

Table 3: Graph density, Diameter, CPL and CL of the DMOZ graphs

Graph	Density	Diameter	CPL	CL
T	0.0002%	14	3.6981	222,286
S	0.0003%	33	7.5791	321,849
R	0.0003%	61	16.2653	4,997
G	0.0006%	45	11.1196	22

4.3.2. Diameter

As explained in the Background section, the diameter of a network is the shortest distance between the two most distant nodes in the network. This measure, jointly with the CPL and the Connectivity Length, provide relevant information about the connectivity patterns of the network. The diameter of the DMOZ graphs are shown in the third column of Table 3. As is expected for the case of T, the diameter is coincident with the depth of the hierarchy. An example of a node corresponding to the path of a topic with that length is "Top/ Regional/ North America/ United States/ New York/ Localities/ N/ New York City/ Brooklyn/ Society and Culture/ Religion/ Christianity/ Catholicism/ Eastern Rites/ Maronite". The component R has the largest diameter value and S exhibits a lower value. Consistently, the diameter of G has an intermediate value due to the set of links contributed by T and S. The large diameter of R could be a consequence of the nature of its edges, given that the

"related" links are not governed by any taxonomical principle as actually are T and S.

4.3.3. Characteristic Path Length and Connectivity Length

The CPL and CL measures correspond to average values associated with the connectedness of the network. The CPL expresses the expected length of any existing path between two nodes in the graph, while the CL provides a more representative estimation by taking into account that the network could contain disconnected components. The fourth and fifth columns of Table 3 show these two measures for each DMOZ graph. While CPL only averages the distances between every pair of co-reachable nodes in the network, the CL penalizes the existence of non-co-reachable nodes, stressing the importance of connectedness in the network. This fact is evidenced for the components T and S that have good (low) CPL values but bad (very high) CL values. The component R has a considerable lower value of CL, but it is still very high to consider that subgraph as highly connected. Evidence of the fact that the *R-subgraph* is not highly connected is also given by its CPL, which takes the highest value among the four analyzed graphs. In contrast, when the three components are considered jointly in G, the CL value decreases to 22, denoting a high connectedness for the complete DMOZ graph. In this case the CPL and CL values are more consistent.

4.3.4. Local clustering coefficient

The Clustering Coefficient (CC) values are shown in table 4 and figure 5. The T component does not contain any cycle, as it is a hierarchy. Also, every node only has edges connecting it with its immediate descendants. Hence, no node can have a CC value different from 0, given that there are no connections between its neighbors.

Figure 5 shows a chart with the grouped frequencies of CC values for each graph. Most values agglomerate in the interval between 0 and 0.1 in all the graphs. As it can be seen in table 4, the highest average CC value corresponds

Table 4: Clustering coefficients values of the DMOZ graphs

Graph	Average CC	Number of nodes with CC=0	Number of nodes with CC=1
T	0	571148	0
S	0.0096	548,312	131
R	0.0484	496,157	2,040
G	0.0531	372,272	935

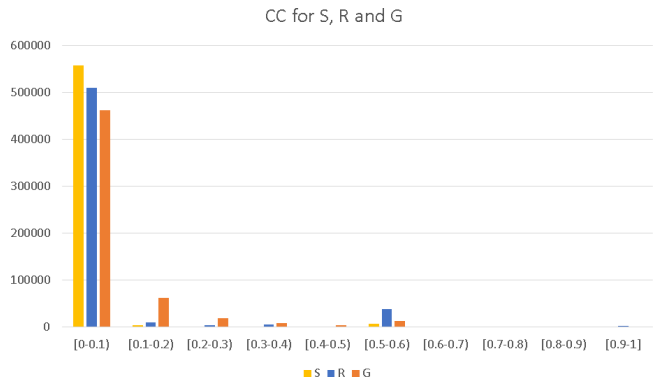

Figure 5: Grouped frequencies for Clustering coefficients of DMOZ graphs

to graph *G*. Nevertheless, it is not a relatively high CC value and the number of nodes with 0 as CC value is very high for all the graphs.

4.3.5. Betweenness centrality

The objective of measures of node centrality is the identification of those nodes that control a network, participating in a large percentage of the communication between its nodes. Table 5 shows the average BC values for the analyzed DMOZ graphs as well as the number of nodes with BC=0 and the maximum BC value found in each graph. Determining the BC measures for the DMOZ graphs is computationally very expensive given that it requires computing the shortest path between all node pairs. As can be seen in table 5, the graph *G* has a node that participates in more than 25 billions of shortest paths,

and the average BC value is about 2.7 millions. Even though there are very high BC values, the proportion of nodes with 0 as BC value is considerably high in each graph. This is a consequence of the taxonomical organization of the *T*- and *S-subgraphs* that position the most specific nodes at the end of the paths, and hence those nodes do not take part as intermediaries in any shortest path. Figure 6 shows the distributions of BC values in each DMOZ graph. According to that figure, while all of these distributions seem to be heavy-tailed, only those associated with the *T*- and *S-subgraph* appear to fit power laws.

Table 5: Betweenness centrality values of the DMOZ graphs

Graph	Average BC	Number of nodes with BC=0	Highest BC
T	22	435,641	529,605
S	44	512,596	963,578
R	25,121	517,718	105,026,375
G	2,708,969	233,362	25,192,567,132

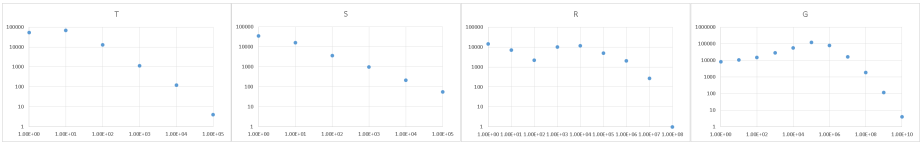

Figure 6: BC distribution in log-log scale for the DMOZ graphs

4.3.6. Alternative Measures of Centrality

Another method for the computation of the centrality of a node in a network consists in determining the set and distances of co-reachable nodes in different ways, instead of considering the shortest paths that include the corresponding node (BC) or its neighbors and their connections (CC). This approach is implemented by the measures of Closeness Centrality, Harmonic Centrality and Lin's Index, described in the Background section. Table 6 shows the mean and max values of these measures for each DMOZ graph.

Table 6: Average values of Closeness Centrality, Harmonic Centrality and Lin's Index for the DMOZ graphs

Graph	Closeness Centrality		Harmonic Centrality		Lin's Index	
	Avg	Max	Avg	Max	Avg	Max
T	0.042939	1	2.57	3.25	1.74	1.87
S	0.256610	1	1.77	589.92	1.51	520.00
R	0.077657	1	114.28	16179.53	109.03	14441.91
G	0.000142	1	25409.22	43242.16	24430.69	39754.73

As can be seen in table 6, all the graphs exhibit a maximum Closeness Centrality value of 1. In particular, the children of the root node of the main taxonomy have a value of 1 for the Closeness Centrality measure, given that the only co-reachable node for them is, in fact, the root node. By definition, the Lin's index of a node i is the product of the Closeness Centrality value and the square of the number of co-reachable nodes of i. Such index corresponds to a weighting scheme for a more accurate estimation of centrality. Furthermore, the Harmonic Centrality tries to capture centrality in a more representative way, taking the harmonic mean of the distances of a node i from its co-reachable nodes. For the case of Closeness Centrality, the highest average value corresponds to the *S-subgraph*, with a 25%, indicating a better mean connectivity for the nodes in that network, according to that measure. However the remaining measures, namely the Harmonic Centrality and Lin's index, point out graph G as the most connected network, and in some cases the difference of the average values is about two orders of magnitude, as seen in table 6. A deeper analysis on the individual nodes could lend more insights into the accuracy of each measure on the estimation of centrality.

4.3.7. Degree Distributions

Scale-free networks are those having degree distributions with power-law behavior, as explained in the Background section. The charts in figure 7 show out- and in-degree distributions in log-log scale for the DMOZ graph and its three subgraphs. For instance, the in-degree distribution plots of the *R-subgraph* and graph G, as well as the out-degrees of the *T-subgraph*, *S-subgraph* and G, reveal a possible power-law behavior. The clear power law distribution that exhibits the in-degree for the *R-subgraph* is probably the result of the arbitrary process of topic association resulting from the editors' criteria. This effect is observed in a lower scale for the in-degree of the *S-subgraph*. For the *R-subgraph*, we can observe the existence of highly disproportionate in-degree values, as is the case of topic "Regional/ North America/ United States/ Arts and Entertainment/ Music" with 2.341 incoming links. Given that the *S-subgraph* represents an alternative taxonomical classification, topic association is guided by a mechanism less arbitrary than for the *R-subgraph*, reflecting the notion of topic inheritance. Note that given the tree structure of the *T-subgraph*, its in-degree is always 1, except for the root node whose indegree is 0.

Regarding the out-degree there is a strong evidence of a power law distribution for the *T-* and *S-subgraphs*. However, the absence of a long tail in the *R-subgraph* rules out a possible power-law behavior for this subgraph.

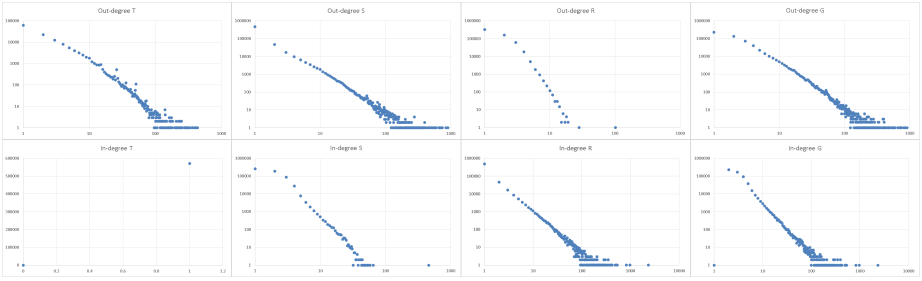

Figure 7: In-degree and Out-degree distributions in log-log scale for the DMOZ graphs

A careful analysis of the out-degree charts of T, S and G in figure 7 reveals a slight curvature, turning the trend line smoothly concave. If this effect is strong

enough, the corresponding distributions could actually depict exponential rather than power-law behaviors [13].

Table 7: Scaling exponents of in-degree, out-degree and degree (for underlying undirected graphs) distributions for the DMOZ graphs

Graph	γ^i	γ^o	γ
T	-	2.26	2.26
S	3.215	2.03	2.132
R	2.012	3.873	2.121
G	2.242	2.265	2.339

The goodness of fit of each dataset to a power-law distribution can be better analyzed in terms of the scaling exponent in each case. The second and third columns of table 7 show the in- and out-degree scaling exponents. Because of the curvature mentioned in the previous paragraph, exponential distributions could fit these models better than power-law distributions, according to Clauset et al. [13]. The highest value observed in the exponents is associated with the out-degree of the *R-subgraph*, and is close to 4, reinforcing the idea of a scale-free behavior for that subgraph. Also the *S-subgraph* has a fairly high scaling exponent for its in-degree.

The fourth column of table 7 shows the scaling exponents of the undirected versions of the DMOZ graphs. The fact that the scaling exponents for the degree distributions of all the analyzed undirected graphs are greater than 2 but smaller than 3 indicates that they have a well-defined mean but do not have a finite variance, as is the case for most recognized power laws in nature. On the other hand, for the particular cases of the in-degree distribution of the *S-subgraph* and the out-degree distribution of the *R-subgraph*, it can be stated that they have a well-defined mean and a finite variance according to the theory of power-law distributions [33], given that their scaling exponents are higher than three.

4.4. Visual analysis of centrality and connectivity

A number of relevant properties can be illustrated or derived from the visualization of the analyzed graphs. Nevertheless, visualizations for depicting certain phenomena are sometimes hard to obtain for the entire graph. The next subsections show selected visualizations that exhibit and reinforce findings derived from the centrality and connectivity analysis carried out in this article. In some cases, the visualizations are presented for the entire graph, while in others, a strategically selected subgraph is visualized to illustrate specific phenomena.

4.4.1. K-Core Decompositions

The LaNet-Vi software tool [2] was employed for rendering panoramic images of the distributions of nodes in the DMOZ graphs, according to the *k-core decomposition algorithm* [38]. As this tool was designed for undirected graphs, all edges in the graphs were transformed to undirected edges. Note that a pair of reciprocal edges in a directed graph becomes a single edge in the corresponding undirected graph. Figure 8 shows the visualizations of the corresponding k-core decompositions for the four DMOZ graphs. As is expected from the tree structure of the *T-subgraph* the value of all the shell degrees (except for the root) is 1. The different lengths of the branches in T have an interesting effect on the k-core visualization. Since the diameter of a shell is a function of the number of nodes belonging to that shell, the large number of leaves in the *T-subgraph* causes a big gap between the outer layer of the k-core decomposition and the remaining layers. Such gap can be also appreciated in the visualizations of *S-* and *R-subgraphs* suggesting, on the one hand, the existence of a great ratio of nodes with very low degree, and on the other hand, a very high difference of degrees between the former set of nodes and the remaining nodes.

In the case of the *S-* and *R-subgraphs*, there are nodes that have shell degrees higher than 1, revealing that these graphs do not have a tree structure, as it was expected. These two graph visualizations exhibit small circles over the circumferences of their outer layers as well as within their displaced central structures and within the gap between the most outer layer and the remaining

more inner layers. These circles correspond to clusters of nodes of the same shell. As such clusters have edges connecting them to nodes of the inner layers, they cannot represent disconnected k-cores, and should have considerably more links to nodes of the outer layers than the rest of the nodes in the inner circles, according to the k-core decomposition algorithm. There are more nodes with high shell degree in the *R-subgraph* visualization, and the shell degrees are slightly higher in that graph than in the *S-subgraph*. The images of the *S-* and *R-subgraphs* exhibit particular filling patterns in their inner layers that correspond to the rendering of a very large number of edges inside the k-cores. The color of each of these edges depends on the k-core of its source and target nodes.

Graph G exhibits well-separated cores, and low overlapping among the cores, suggesting a regular structure in the connections between nodes. It is also remarkable the existence of small circles outside the border of the central core, that could imply the existence of smaller communities of topics with very few connections to the nodes of that core.

4.4.2. Strongly Connected Components of G

Three strongly connected components of the graph G are visually represented in figure 9, using the *Gephi* and *LaNet-Vi* software tools. For the Gephi visualization of the largest SCC of G (bottom left) labels are displayed only for a small number of nodes with very high degree. The purpose of a visual analysis of a SCC is to identify connectivity patterns that give place to a characterization of the corresponding subgraph. Another feature that can be observed through the visualizations in figure 9 is the strong cohesion among the topics included in some SCC's. Regarding this fact, it is important to state that all the nodes in the first two SCC's are highly related to one another, within each SCC. The nodes of the component at the top-left corner of the figure are all descendant of topic "Games/ Video Games", and refer to cheats and hints of video-games of different genres and platforms. Besides, the nodes of the component at the middle-left part are related to descendants of "World/ Italiano"

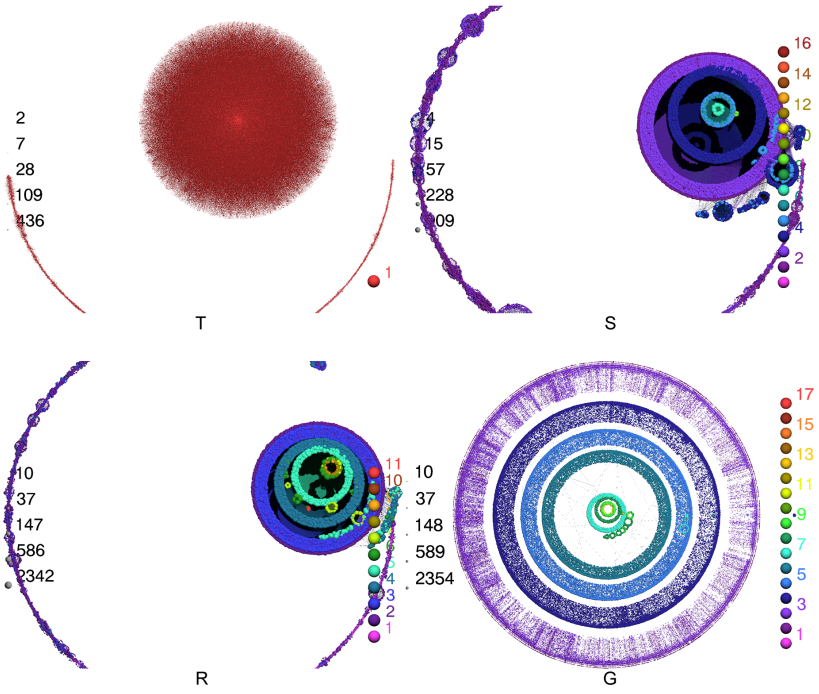

Figure 8: K-core decomposition visualizations of the DMOZ graphs

that have the name "Immobiliari" and correspond to the topic "Real-Estate" in different regions of Italy. Finally, the nodes of the component at the bottom-left corner belong to the greatest SCC found in G. The "Video-games" component consists of 128 nodes, the "Real-estate" component consists of 122 nodes and the greatest SCC is composed by 267,541 nodes. According to the Gephi and LaNet-Vi visualizations, the maximum degree found among the nodes in the "Video-games" SCC is 57 (68 for the directed graph) and is associated with the most representative topic in that component, namely "Top/Games/Video Games/Cheats and Hints". For the case of the "Real-estate" SCC, the highest degree is 21 (40 for the directed graph), corresponding to the main topic "World/ Italiano/ Regionale/ Europa/ Italia/ Affari e Economia/ Immobiliari". In contrast to this, the greatest SCC of G has a highest degree of 2,353 (2,356 for the

directed graph). The highest shell degree values for the k-core decompositions are 4, 3 and 12 for each of the three analyzed SCC's. The first two components are two samples of cohesive groups of nodes arising from the process of SCC decomposition. Even though a deeper analysis is required, this could imply that meaningful relatively small topical communities can be found by applying the SCC detection algorithm. On the other hand, for the largest SCC it is possible to identify topics that are not semantically related, as is the case for "Music", "Windows", "Business and Economy", etc.

4.5. Connectivity and Centrality of some relevant topics

This section offers a focalized analysis of DMOZ by identifying specific topics that consistently present high centrality and connectivity values for the analyzed DMOZ graphs. Table 8 specifies the graphs where these topics exhibit the highest values and reports their associated centrality and connectivity measures, including degree, betweenness centrality, closeness centrality, harmonic centrality and Lin's index.

A comprehensive analysis of the clustering coefficient (CC) has revealed that there is a great number of nodes achieving the maximum value for each graph. Besides, these specific nodes do not exhibit relevant values for the remaining measures, and the nodes of table 8 do not have particularly high CC values. As a consequence, the clustering coefficient column is omitted in this table.

5. Discussion

The analysis presented in this article provides insight about several aspects of the DMOZ topic ontology. Among the most important aspects in modeling a network are the communication patterns between its actors. It is well known that small-world networks [41] exhibit good communication capabilities. For the case of DMOZ, good communication patterns imply ease of navigability. As mentioned earlier, different analyses in cognitive science have revealed that semantic networks of words and concepts exhibit the small-word property

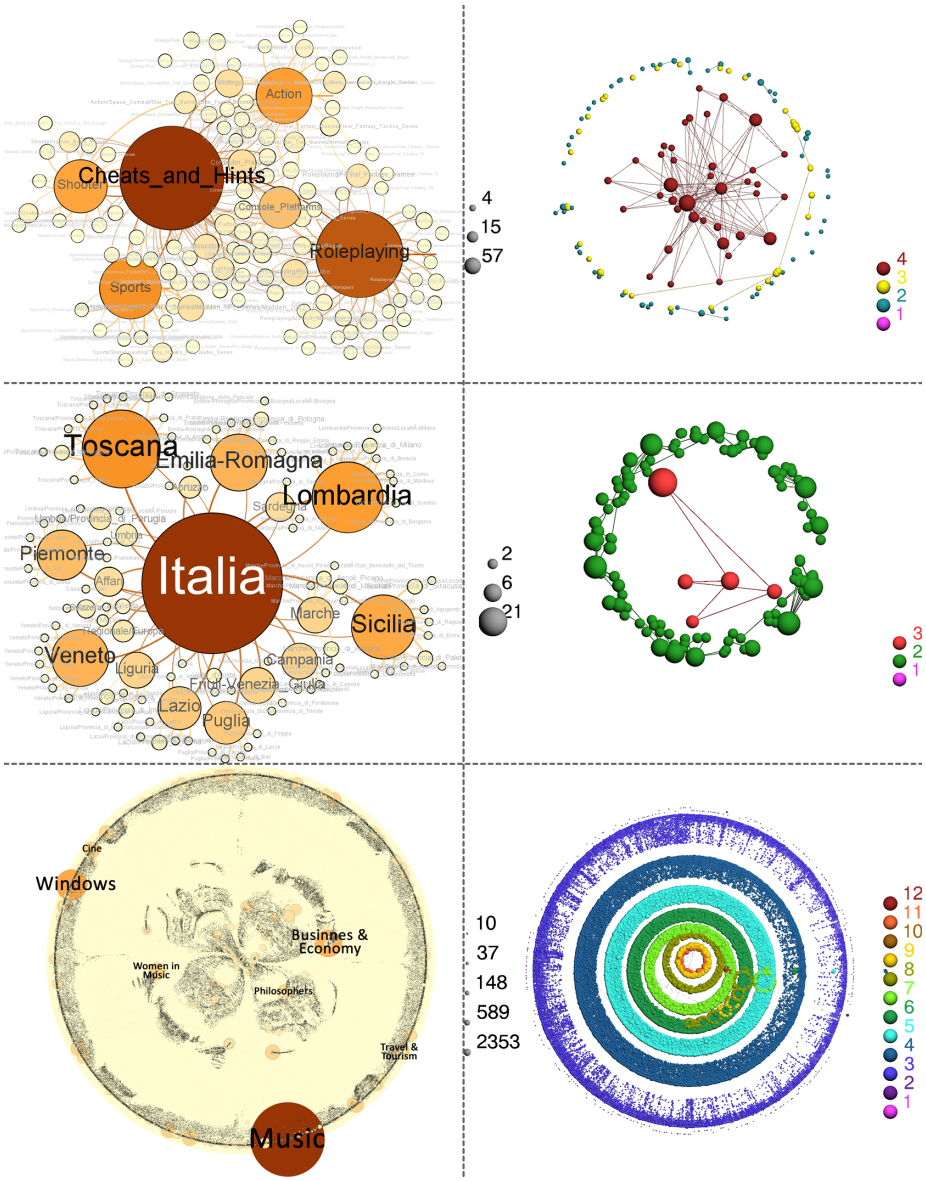

Figure 9: Gephi (left) and LaNet-Vi (right) visualizations of three SCC's of G. At the top, the SCC corresponding to the topic "Video Games". At the middle, the SCC corresponding to the topic "Immobiliari" (Real-estate) in Italy. At the bottom, the greatest SCC of G

([39, 7, 31]), which implies that words and concepts tend to form semantically cohesive communities, and most words and concepts are typically at a short

Table 8: Most central topics of DMOZ graphs. The rows correspond to the topics. The columns are associated with the connectivity and centrality measures considered. Each cell of the table enumerates the graphs where the corresponding topic has the highest value.

Topic	Degree	Betweenness centrality	Closeness centrality	Harmonic centrality	Lin's index
Regional		G			
Adult			G		
Science/ Environment		R		R	R
World/ Español/ Artes/ Cine	S			S	S
Regional/ North America/ United States/ Arts and Entertainment/ Music	R, G				

semantic distance between each other. We argue that these phenomena should also manifest themselves in well-specified topic networks since it is natural for topics and subtopics to form thematically coherent clusters. In addition, it is commonly agreed that "all knowledge is connected to all knowledge" and therefore it is expected that most topics be directly or indirectly related through short meaningful paths. As a consequence, we state that a good network topology for a topic ontology should be one in which nodes tend to cluster together (high Average Clustering Coefficient) and in the meantime most nodes should be reachable from any other node in a few steps (small CL and CPL), providing good navigability among topics.

The scale-free structure seems to also be a feature of the analyzed DMOZ graphs. The degree distributions shown in figure 7 and the scaling exponent

values for the in- and out-degree distributions as well as for the degree distributions of the undirected graphs presented in table 7 indicate that all these data fit power laws. However, studies on the process of formation and evolution of the underlying networks are required to draw conclusions on related concepts such as network growth and preferential attachment [3].

There are some nodes that seem to be consistently central under different graphs and metrics as seen in table 8. Such is the case of the general topics "Regional" and "Adult". This consistency is observed also for the topic "Science/ Environment" and highly specific topics, i.e. topics placed in deep levels of the taxonomy, as is the case of "World/ Español/ Artes/ Cine" and "Regional/ North America/ United States/ Arts and Entertainment/ Music". This fact can be the effect of the multiplicity of links from other topics, which may be due to different classifying criteria employed by the editors of the DMOZ directory. Regarding topic "Regional/ North America/ United States/ Arts and Entertainment/ Music", it is pointed out as the node with highest degree both in G and in the *R-subgraph*. This is due to the high number of incoming links to this node in the *R-subgraph*. Also, this is the most connected node of the greatest SCC of G, as seen in figure 9.

6. Conclusions and Future work

This article described the application of complex network theory to the study of the DMOZ topic ontology. The set of analyses presented offer a characterization of the most important topological aspects underlying the large-scale structure of DMOZ. According to the results obtained from these analyses, the largest human-edited topic ontology in the web exhibits a number of interesting non-trivial regularities also found in other natural and artificial complex networks.

The main regularities encountered in DMOZ are the small-world and scale-free phenomena. Despite the power-law behavior exhibited by the degree frequency histograms of the DMOZ graphs, according to Clauset et al. [13] further

statistical analyses should be carried on to determine whether other statistical distributions fit better the corresponding networks. The scale-free behavior is not only reflected in the degree distribution but also in the power-law distribution of SCC sizes.

Scale-free networks are often dynamically growing networks. Evidence about the temporal dynamics of the DMOZ network still needs to be gathered and investigated. In order to characterize the evolution process of the associated semantic network, further algorithmic models should be designed and simulated, with knowledge about the relative age of each node in the DMOZ graph. An approach similar to the one described by Steyvers et al. [39] could be employed for this analysis, considering possible indicators of the time when each topic was added to the taxonomy. Such studies could also be linked and contrasted with the semantic networks of subject memories analyzed by Morais et al. [31]. Besides, the evolution process of the DMOZ ontology should be analyzed in order to determine whether a preferential attachment mechanism occurs favoring the oldest topics in the directory, as described by Downes [17] regarding the power-law nature of connective knowledge.

Some interesting conclusions emerge in relation to the detection of SCC's in the DMOZ graphs. In the first place, as seen in the Visual analysis section, the small SCC's exhibit cohesive topical communities. The existence of a Great SCC suggests further studies, including calculations of specific measures as the *Percolation Threshold*, intended as the number of nodes that should be removed from the network in a targeted attack for causing the disappearance of the Great SCC. A similar study has been carried out by De La Torre et al. [16].

Different from many studies on complex networks, the analysis performed here is not carried out on a sample but on the totality of the DMOZ network, avoiding possible artifacts that may result from incompleteness or sampling biases. In spite of that, some artifacts are still observable, which we argue may result from existing constraints on the ontology structure, such as its underlying taxonomic hierarchy.

As stated by Resnik [37] and further discussed by Xamena et al. [42], while

topological measures can provide useful insights into the similarity and relevance relation existing between concepts or topics, proximity is sometimes ineffective as an estimator of the semantic relation between objects in certain classes of networks where links do not represent uniform distances. An analysis of a network that is based only on topological aspects and disregards a deeper analysis of the content of the nodes and the semantic of the relations seems to be insufficient to derive meaningful measures of topic relevance and semantic similarity. As part of our future work we plan to investigate if the combination of topological features and content-based features of DMOZ can be used to infer good predictors of topical relevance and similarity. This kind of relations are observed in the work of Mika [30], where it is stated that the most general concepts often present low Clustering coefficients and high BC values, and the most specific concepts exhibit exactly the opposite behavior in semantic networks. Another interesting ingredient in the work of Mika [30] that could be employed in DMOZ is the inclusion of the social component in semantic networks by means of the widely known "folksonomies". Those structures can derive better semantic schemes in particular types of social contexts.

Finally, in order to gather additional insight into the structure of DMOZ, we plan to analyze whether correlations exist between some of the measures reported in this study, or between these measures and the number of websites indexed under the topics. We also plan to investigate whether correlations between nodes of similar degree exist with the purpose of measuring assortativity, which is a tendency commonly observed in some networks where nodes tend to attach to other nodes that are similar in some way.

Acknowledgments

This work was supported by CONICET (PIP 11220120100487, PIP 11220120100309CO), MinCyT (PICT 2014-0624, PICT 2012-0691), and Universidad Nacional del Sur (PGI-UNS 24/N029, PGI-UNS 24/N034).

References

[1] L. A. Adamic, B. A. Huberman, Power-law distribution of the world wide web, Science 28 (5461) (2000) 2115–2115.

[2] J. I. Alvarez-Hamelin, L. Dall'Asta, A. Barrat, A. Vespignani, Large scale networks fingerprinting and visualization using the k-core decomposition, in: Advances in neural information processing systems, 2005, pp. 41–50.

[3] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (5439) (1999) 509–512.

[4] M. Bastian, S. Heymann, M. Jacomy, et al., Gephi: an open source software for exploring and manipulating networks., ICWSM 8 (2009) 361–362.

[5] A. Bavelas, Communication patterns in task-oriented groups., Journal of the acoustical society of America.

[6] P. Boldi, S. Vigna, Axioms for centrality, Internet Mathematics 10 (3-4) (2014) 222–262.

[7] J. Borge-Holthoefer, A. Arenas, Semantic networks: structure and dynamics, Entropy 12 (5) (2010) 1264–1302.

[8] U. Brandes, A faster algorithm for betweenness centrality*, Journal of mathematical sociology 25 (2) (2001) 163–177.

[9] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener, Graph structure in the web, Computer networks 33 (1) (2000) 309–320.

[10] E. Bullmore, O. Sporns, Complex brain networks: graph theoretical analysis of structural and functional systems, Nature Reviews Neuroscience 10 (3) (2009) 186–198.

[11] E. Bullmore, O. Sporns, The economy of brain network organization, Nature Reviews Neuroscience 13 (5) (2012) 336–349.

[12] C. Cattuto, C. Schmitz, A. Baldassarri, V. D. Servedio, V. Loreto, A. Hotho, M. Grahl, G. Stumme, Network properties of folksonomies, Ai Communications 20 (4) (2007) 245–262.

[13] A. Clauset, C. R. Shalizi, M. E. Newman, Power-law distributions in empirical data, SIAM review 51 (4) (2009) 661–703.

[14] T. F. Coleman, J. J. Moré, Estimation of sparse jacobian matrices and graph coloring blems, SIAM journal on Numerical Analysis 20 (1) (1983) 187–209.

[15] L. d. F. Costa, O. N. Oliveira Jr, G. Travieso, F. A. Rodrigues, P. R. Villas Boas, L. Antiqueira, M. P. Viana, L. E. Correa Rocha, Analyzing and modeling real-world phenomena with complex networks: a survey of applications, Advances in Physics 60 (3) (2011) 329–412.

[16] S. R. de la Torre, J. Kalda, R. Kitt, J. Engelbrecht, On the topologic structure of economic complex networks: Empirical evidence from large scale payment network of estonia, Chaos, Solitons & Fractals 90 (2016) 18–27.

[17] S. Downes, An introduction to connective knowledge, http://www.downes.ca/post/33034 (2005).

[18] V. M. Eguiluz, D. R. Chialvo, G. A. Cecchi, M. Baliki, A. V. Apkarian, Scale-free brain functional networks, Physical review letters 94 (1) (2005) 018102.

[19] L. C. Freeman, A set of measures of centrality based on betweenness, Sociometry (1977) 35–41.

[20] A. Gibbons, Algorithmic graph theory, Cambridge University Press, 1985.

[21] C. Guéret, P. Groth, C. Stadler, J. Lehmann, Assessing linked data mappings using network measures, in: Extended Semantic Web Conference, Springer, 2012, pp. 87–102.

[22] F. Harary, et al., Graph theory, Reading: Addison-Wesley.

[23] B. Hoser, A. Hotho, R. Jäschke, C. Schmitz, G. Stumme, Semantic network analysis of ontologies, in: European Semantic Web Conference, Springer, 2006, pp. 514–529.

[24] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, A.-L. Barabási, The large-scale organization of metabolic networks, Nature 407 (6804) (2000) 651–654.

[25] J. Kleinberg, The small-world phenomenon and decentralized search, SiAM News 37 (3) (2004) 1–2.

[26] L. F. Lago-Fernández, R. Huerta, F. Corbacho, J. A. Sigüenza, Fast response and temporal coherent oscillations in small-world networks, Physical Review Letters 84 (12) (2000) 2758.

[27] N. Lin, Foundations of social research, McGraw-Hill New York, 1976.

[28] A. G. Maguitman, F. Menczer, F. Erdinc, H. Roinestad, A. Vespignani, Algorithmic computation and approximation of semantic similarity, World Wide Web 9 (4) (2006) 431–456.

[29] M. Marchiori, V. Latora, Harmony in the small-world, Physica A: Statistical Mechanics and its Applications 285 (3) (2000) 539–546.

[30] P. Mika, Ontologies are us: A unified model of social networks and semantics, Web semantics: science, services and agents on the World Wide Web 5 (1) (2007) 5–15.

[31] A. S. Morais, H. Olsson, L. J. Schooler, Mapping the structure of semantic memory, Cognitive Science 37 (1) (2013) 125–145.

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

[33] M. E. Newman, Power laws, pareto distributions and zipf's law, Contemporary physics 46 (5) (2005) 323–351.

[34] S. Nikolova, J. Boyd-Graber, C. Fellbaum, Collecting semantic similarity ratings to connect concepts in assistive communication tools, in: Modeling, Learning, and Processing of Text Technological Data Structures, Springer, 2011, pp. 81–93.

[35] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, A. Vespignani, Epidemic processes in complex networks, Reviews of modern physics 87 (3) (2015) 925.

[36] R. Pastor-Satorras, A. Vespignani, Evolution and structure of the Internet: A statistical physics approach, Cambridge University Press, 2007.

[37] P. Resnik, Using information content to evaluate semantic similarity in a taxonomy, in: Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1, IJCAI'95, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1995, pp. 448–453.

[38] S. B. Seidman, Network structure and minimum degree, Social networks 5 (3) (1983) 269–287.

[39] M. Steyvers, J. B. Tenenbaum, The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth, Cognitive science 29 (1) (2005) 41–78.

[40] R. Tarjan, Depth-first search and linear graph algorithms, SIAM journal on computing 1 (2) (1972) 146–160.

[41] D. J. Watts, S. H. Strogatz, Collective dynamics of "small-world" networks, nature 393 (6684) (1998) 440–442.

[42] E. Xamena, N. B. Brignole, A. G. Maguitman, A study of relevance propagation in large topic ontologies, Journal of the American Society for Information Science and Technology 64 (11) (2013) 2238–2255.