Structural analysis of relevance propagation models

Article *in* Knowledge-Based Systems · October 2021

DOI: 10.1016/j.knosys.2021.107563

3 authors:

Eduardo Xamena
National Scientific and Technical Research Council
9 PUBLICATIONS **7** CITATIONS
SEE PROFILE

Nélida B. Brignole
National Scientific and Technical Research Council
84 PUBLICATIONS **395** CITATIONS
SEE PROFILE

Ana Gabriela Maguitman
Universidad Nacional del Sur
120 PUBLICATIONS **2,302** CITATIONS
SEE PROFILE

Structural Analysis of Relevance Propagation Models

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

[a]*Instituto de Investigaciones en Ciencias Sociales y Humanidades (ICSOH) - CONICET - UNSa; Departamento de Informática - Facultad de Ciencias Exactas - Universidad Nacional de Salta (UNSa) - Av. Bolivia 5150, Salta, Argentina.*
[b]*LIDeCC - Laboratorio de Investigación y Desarrollo en Computación Científica - San Andrés, Bahía Blanca, Argentina.*
[c]*Instituto de Ciencias e Ingeniería de la Computación (ICIC), CONICET, UNS, San Andrés 800, Campus Palihue, Bahía Blanca, Argentina.*
[d]*DCIC-UNS - Departamento de Ciencias e Ingeniería de la Computación - Universidad Nacional del Sur - San Andrés 800, Bahía Blanca, Argentina.*
[e]*LIDeCC - Laboratorio de Investigación y Desarrollo en Computación Científica (DCIC-UNS); Planta Piloto de Ingeniería Química (UNS-CONICET), Cno la Carrindanga km 7, Bahía Blanca, Argentina.*

Abstract

Relevance relations constitute the core of information retrieval. Topical ontologies, such as collaborative webpage classification projects, can provide a basis for identifying and analyzing such relations. New meaningful relevance relations can be automatically inferred from these ontologies by composing existing ones. In this work, several relevance propagation models are analyzed in terms of complex network theory. Structural properties such as Characteristic path length, Clustering coefficient and Degree distribution are computed over the models in order to understand the nature of each underlying network. This analysis raises interesting points about the Small-world and Scale-free structure of some relevance propagation models. Moreover, other connectivity and centrality measures are computed to gain additional insight into the topology of relevance. Finally, the analysis is complemented by providing visualizations of the k-core decomposition of different relevance propagation models. To illustrate the generalizability of the proposed methodology the analysis is carried out on an ontology from a different domain. The major theoretical implication

Corresponding author email: examena@di.unsa.edu.ar

of this analysis is the derivation of new instruments to typify semantic networks derived from relevance relations. The results can be exploited in a pragmatic way, as the parameters and properties derived by this analysis can serve as prior knowledge to algorithms for the automatic or semi-automatic construction of semantic networks.
Keywords: relevance propagation, topic ontologies, complex networks, topological analysis

---

1. Introduction

The notion of relevance is crucial to information retrieval. Ontologies are commonly used in information retrieval tasks as they are designed to organize knowledge, helping to capture relevance relations in a simplified manner. DMOZ[1] is a web directory and a special ontology, built collaboratively by users around the world. The latest version of DMOZ comprises more than 1,000,000 topics arranged hierarchically by taxonomical edges, and complemented with two different kinds of cross edges, namely "symbolic" and "related" links. Complex network theory provides a comprehensive set of tools for analyzing large graph representations. In general, the strength and robustness of a network can be evaluated in terms of metrics that rely on simple concepts, as the number of connections of a node, the number of connections among the neighbors of a node or the number of paths that could be established between two nodes. Several works employ these measures and features to characterize different networks, as is the case of brain networks [11], peer-to-peer systems [42], characters in fictional novels [30], and the entire web [10]. As ontologies are frequently represented by large graphs, complex network theory provides useful tools for their analysis. New topologies emerge when a basic ontology graph is augmented with implicit relations, and topological measures allow to characterize these augmented models.

[1]The DMOZ archive is hosted at `http://dmoz-odp.org/` and is continued by Curlie at `https://curlie.org/`.

Relevance has traditionally been studied from a logical or philosophical perspective (e.g., [6]). Also, for years, the notion of relevance has been addressed from an information retrieval perspective (e.g., [18]). However, to the best of the authors' knowledge, relevance has never been analyzed from a topological perspective.

Our analysis focuses on topological regularities observed in different Relevance Propagation Models (RPM's) proposed in [52]. In that work several RPM's were analyzed qualitatively and quantitatively, evaluating their accuracy in terms of human users' criteria. Later, a study of the DMOZ topology was presented in [53]. However, this study was limited to the basic DMOZ graph, without considering relevance propagation. The present work extends the results reported in [52, 53] by examining various complex network properties of the resulting RPM's, such as node degree distributions, local clustering coefficient, average shortest path length, diameter of the network, among other specialized metrics. This analysis allows to identify non-trivial regularities and provides the basis for a better understanding of important notions such as connectivity, topic relevance and semantic similarity, among others. This novel approach provides a different view on the notion of topical relevance, allowing to study the multitude of interactions that occur between topics. It offers a new perspective to analyze computational models that seek to explain how relevance propagates and how semantic structures form.

The analysis presented in this article provides new tools to the active research area of semantic network construction [12] or expansion [55]. As is the case with other complex networks, the description of topological properties of semantic networks such as ontologies or knowledge graphs can be done using power-laws and other metrics observed in this kind of networks. As a consequence, instead of using average values to characterize different aspects of the graph, exponents of power laws and other relevant characteristics derived from a topological analysis as the one presented here can be used to typify the statistical aspects of the graph. Also, our analysis can guide the design of these networks or the incorporation of new elements to an existing one. The charac-

teristics identified in our analysis can help the process of enriching an ontology or knowledge graph with upcoming data, while at the mean time preserving several statistical regularities that are observed in any relevance propagation model.

Manually building large ontologies, knowledge graphs or other semantic networks (as the one analyzed here) can be a tedious task. We contend that our analysis can provide valuable parameters that can be exploited by structure learning algorithms for the automatic or semi-automatic generation of these networks [1]. It is worth mentioning that many structure learning algorithms assume or require prior knowledge on the distribution associated with the structure of the graph to be inferred. Hence, we anticipate that any algorithm that attempts to build a semantic network from automatic relevance relation extraction can highly benefit from prior knowledge on structural aspects of the "network of relevance relations" to be derived. As pointed out by other studies [48] "structure always affects function". Hence, a good relevance propagation model may have to be sensitive to its underlying structure. As a consequence, the topological analysis reported in this work serves as a key complement to the comparative studies of RPM's presented in [52].

This article is organized as follows. Section 2 describes the DMOZ graph, RPM's and other related concepts to provide a basis for the remaining sections. Then, section 3 reviews literature on structural analysis, semantic networks and other related topics. The main results of this work are reported in section 4. In the first place, this section provides an overview of the datasets used in our analysis and the computed RPM's. It then investigates and discusses several properties of the DMOZ graph and its derived RPM's by means of metrics commonly used in complex network analysis. In particular, it analyzes the connectivity patterns of the networks, reports centrality measures, explains the Small-world and Scale-free properties on the RPM's and provides visualizations of the graphs' k-core decomposition and the in- and out-degree distributions. Section 5 investigates the generalizability of the proposed methodology by providing a brief analysis on an alternative ontology. Finally, the conclusions and

possible future research avenues are presented in section 6.

2. Background

In this section we present the notion of topic ontology and describe the basic structure of the DMOZ graph as a special case of topic ontology. Furthermore, the notion of relevance propagation is reviewed and illustrated with an example. Finally, we review measures and tools from complex network theory employed in the analysis presented in this article.

2.1. Topic Ontologies and the DMOZ graph

A topic ontology is a set of topics connected with different types of relations. Each topic includes a set of related documents. A well-known example of topic ontology is the DMOZ project, which is a collaborative effort for classifying websites into a topical structure. Such a structure results in a big graph or ontology that contains three kinds of links:

- T: "is-a" links, representing the hierarchical component of the ontology,
- S: non-hierarchical "symbolic" cross links,
- R: "related" cross links, also organized in a non-hierarchical manner.

While the hierarchical component imposes strong constraints on the general organization of the DMOZ ontology, the "symbolic" and "see-also" connections loose up these constraints and offer the possibility of integrating the taxonomical component of DMOZ with more general components. The DMOZ ontology can be formally characterized as a graph G = (V, E) with a set of vertices V representing topics and a set of edges E representing the union of the three types of links (i.e., $E = T \cup S \cup R$). A portion of this graph is illustrated in Figure 1.

2.2. Relevance Propagation Models

The notion of relevance in the context of a web directory such as DMOZ has been defined and discussed in [52]. According to that work, a topic t_i is relevant

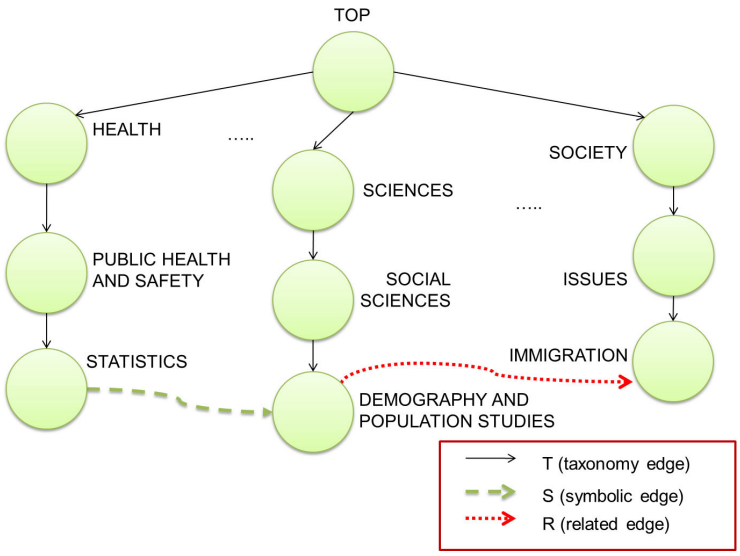

Figure 1: Portion of the DMOZ ontology.

to another topic t_j if there is an edge of some type from topic t_i to topic t_j in the graph representation of the directory. For instance, in Figure 1, we can observe an explicit relevance relation between topics "*Top / Science / Social Sciences / Demography and Population Studies*" and "*Top / Society / Issues / Immigration*", represented by the red dotted arrow that depicts a "related" link. The same applies to other explicit relevance relations represented by taxonomic and symbolic links.

In addition to the relevance relations that are explicitly represented in the graph, a great number of meaningful relevance relations can be identified by taking into account paths between certain topics. An example of a meaningful transitive relation emerges between "*Top / Science / Social Sciences*" and "*Top / Society / Issues / Immigration*". This relation is derived from the path leading from the first topic to the second one by combining taxonomical and "related" edges.

Figure 2 shows the graphical notation employed to describe non-basic RPM's. The source node is represented by a blank circle and the target node by a filled circle. Arrows represent different types of edges and may stand for explicit or

implicit relevance relations. This figure illustrates how implicit relevance relations can be derived from explicit ones. In this case, a new edge e_{ts} results from composing a symbolic edge e_s with a taxonomy edge e_t.

RPM's can be sparser or denser depending on the criteria applied to derive implicit relations. Usually, there is a trade-off between the density of an RPM and the significance of its relevance relations: As more links are established between topics, the real significance of the arising relations tends to decay. While RPM's can be derived in a wide variety of ways from the DMOZ ontology, this article focuses on studying the RPM's described in the Analysis section.

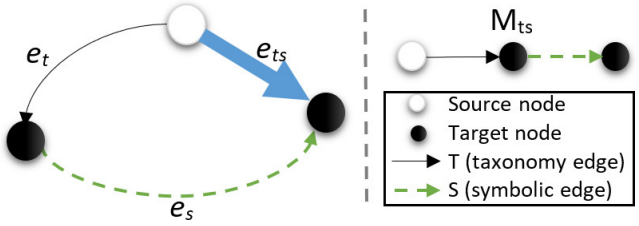

Figure 2: Left-hand side: The composition of edges e_t (taxonomical) and e_s (symbolic) gives rise to an implicit edge e_{ts}. Right-hand side: Graphical expression of the arising RPM.

2.3. Matrix Representation of Relevance Relations

Relevance relations can be represented by means of Boolean matrices. Each row or column of a relevance matrix represents a DMOZ topic, and each cell contains a 1 if there exists a relevance relation between the corresponding row and column topics, and 0 otherwise. These matrix representations facilitate the computation of new relevance relations by means of two basic operations:

- **Union:** Given the relations ρ_A and ρ_B, their union, $\rho_A \cup \rho_B$ is computed as: $[\mathbf{A} \vee \mathbf{B}]$, where $\mathbf{A}$ and $\mathbf{B}$ are the matrix representations of ρ_A and ρ_B, respectively, and the Boolean addition (disjunction) operation $\vee$ on matrices is defined as $[\mathbf{A} \vee \mathbf{B}]_{ij} = \mathbf{A}_{ij} \vee \mathbf{B}_{ij}$.

- **Composition:** Given two relations ρ_A and ρ_B, the composition $\rho_A \circ \rho_B$ can be computed as: $[\mathbf{A} \otimes \mathbf{B}]$, where $\mathbf{A}$ and $\mathbf{B}$ are the matrix represen-

tations of ρ_A and ρ_B, respectively, and the Boolean product operation $\otimes$ on matrices is defined as $[\mathbf{A} \otimes \mathbf{B}]_{ij} = \bigvee_k (\mathbf{A}_{ik} \wedge \mathbf{B}_{kj})$.

The Boolean product can be repeatedly applied on a relevance matrix to compute new matrices representing paths of specific lengths. Let $\mathbf{I}$ be the identity matrix and let $\mathbf{M}$ be a relevance matrix. Then, for any non-negative integer $\mathbf{n}$, $\mathbf{M}^{(\mathbf{n})}$ represents the relevance matrix derived from $\mathbf{M}$ containing paths of length $\mathbf{n}$, and is computed as follows: $\mathbf{M}^{(0)} = \mathbf{I}$, and $\mathbf{M}^{(r+1)} = \mathbf{M} \otimes \mathbf{M}^{(r)}$

Finally, the reflexive-transitive closure of $\mathbf{M}$ is computed as follows:

$$\mathbf{M}^* = \bigvee_{r=0}^{\infty} \mathbf{M}^{(r)} \tag{1}$$

Matrix $\mathbf{M}^*$ codifies whether a topic is reachable from another.

2.4. Complex Network analysis

This section reviews the concepts and measures of complex network theory that we have adopted to analyze the most salient properties of the DMOZ graph and its associated RPM's. The main definitions can be found in this section while a more detailed description and the motivations for each measure are available as supplementary material.[2]

2.4.1. Connectivity and centrality measures

We describe next the connectivity and centrality measures used in this work, for a generic graph $G = (V, E)$, and considering the distance between nodes i and j, $d(i, j)$, as the shortest path length between these nodes.

- **Graph density**:

$$Density(G) = \frac{|E|}{|V|(|V| - 1)} \tag{2}$$

- **Diameter**: The largest distance between any pair of connected nodes, considering distance as the length of the shortest path.

[2] `https://github.com/edus1984/Structural-Analysis-of-Relevance-Propagation-Models`

- **Characteristic (Average) Path Length (*CPL*)**:

$$l(G) = \frac{1}{|V| * (|V| - 1)} \sum_{i \in V} \sum_{j \in V \setminus \{i\}} d(i, j) \tag{3}$$

- **Connectivity length (*CL*)**:

$$D(G) = \frac{|V|(|V| - 1)}{\sum_{i,j \in G} \frac{1}{d(i,j)}}, \tag{4}$$

where we assume the distance between unreachable nodes is ∞ and $\infty^{-1} = 0$.

- **Local Clustering Coefficient (for a node i)** :

$$C_i = \frac{|\{(j, k) : j, k \in N_i; (j, k) \in E\}|}{|N_i|(|N_i| - 1)}, \tag{5}$$

where

  - N_i is the set of neighbors of node i
  - (j, k) is an edge between nodes j and k

We use CC to represent the average clustering coefficient of a network.

- **Betweenness Centrality**:

$$bc_i = \sum_{u \neq i \neq v} \frac{\sigma_{uv}(i)}{\sigma_{uv}}, \tag{6}$$

where:

  - σ_{uv} is the total number of shortest paths from a node u to another node v
  - $\sigma_{uv}(i)$ is the number of shortest paths from node u to node v that pass through i

- **Closeness Centrality**:

$$cl_i = \frac{1}{\sum_{d(j,i) < \infty} d(j, i)} \tag{7}$$

- **Harmonic Centrality**:

$$hc_i = \sum_{d(j,i)<\infty, j\neq i} \frac{1}{d(j,i)} \tag{8}$$

- **Lin's index for Closeness Centrality**:

$$lin_i = \frac{|\{j : d(j,i) < \infty\}|^2}{\sum_{d(j,i)<\infty, j\neq i} d(j,i)} \tag{9}$$

2.4.2. Degree distribution

The degree of a node i in an undirected graph is the number of edges between i and other nodes. In the case of directed graphs, **in-degree(i)** is the number of edges from other nodes to i, while **out-degree(i)** is the number of edges from i to other nodes. The *degree distribution* $\mathrm{P}(m)$ 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* $\mathrm{P}_{in}(m)$ and the *out-degree distribution* $\mathrm{P}_{out}(m)$ are defined as the probabilities for any node of having m incoming or outgoing links respectively:

$$\mathrm{P}_{in}(m) = \frac{\text{Number of nodes with in-degree m}}{|V|} \tag{10}$$

$$\mathrm{P}_{out}(m) = \frac{\text{Number of nodes with out-degree m}}{|V|} \tag{11}$$

2.4.3. "Small World" networks

A *Small-World* network [51, 33] exhibits a low *CPL* $l(G)$ and high levels of *clustering coefficient*. These characteristics turn any node reachable in relatively few steps from any other node (low *CPL*), and result in high connectivity for the entire graph (high clustering coefficient).

2.4.4. Power-law distributions

A Power-law distribution can be modeled with the following probability distribution function [35]:

$$\mathrm{P}(x) = Cx^{-\alpha} \tag{12}$$

Networks that exhibit a power-law degree distribution are said to be scale-free.

3. Related work

This section reviews related work. In particular, it presents an overview of literature on complex network theory applied to semantic networks and outlines some complementary literature on Relevance and Relevance propagation.

Semantic networks are a graphical notation for representing knowledge as a set of interconnected conceptual entities. They have been applied to represent data, reveal structure and support navigation, and they have been widely studied both from the cognitive-science and knowledge-engineering perspectives. On the one hand, the cognitive-science approach to the study of semantic networks focuses on understanding how knowledge is organized in the human brain. On the other hand, the knowledge-engineering perspective puts the emphasis on analyzing semantic networks built by humans, either individually or collaboratively.

The works of Steyvers & Tenenbaum [46], Morais et al. [34] and Mak & Twitchell [32] constitute examples of the study of semantic networks as complex networks from the cognitive-science perspective. According to these authors, complex network theory provides useful tools for the analysis of the organization of concepts in the human brain. Similarly, the study of how the brain networks are structurally organized has been an important area of research in cognitive linguistic, psychology and neuroscience. 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 [11, 45].

More closely related to our work is the study of semantic networks built by humans, such as ontologies or knowledge graphs. These semantic networks are typically built directly by domain experts or elicited from them by knowledge engineers. An early example of the analysis of this kind of networks from the complex network perspective is presented in [21], which focuses on examining the ontological part of the Semantic Web. The studied network contains 307,231 nodes and 588,890 arcs and is the result of combining 282 ontologies

collected from the DAML Ontology Library[3]. Based on metrics and properties of the studied network, such as clustering coefficient, average path length and degree distribution, the authors conclude that the Semantic Web behaves like a complex system. The analysis is also completed for smaller ontologies, yielding the conclusion that the same properties can be observed at different scales. A similar analysis is carried out in [50], where the authors also conclude that the analyzed ontologies reveal several patterns and regularities typically found in complex networks. Also, they claim that these ontologies contain a few "focal" classes that form the conceptual backbone of the defined schema, and many "peripheral" ones that provide details on the former. Ontologies have also been studied as complex networks with the goal of helping ontology engineers understand their structural complexity and evaluate their quality with respect to modular design principles [43].

Hoser et al. [27] illustrate the benefits of computing centrality and connectivity metrics on ontologies to gain insight into the importance of certain concepts and properties of the ontology. Following a similar premise, Zhang et al. [57] propose an approach for automatic ontology summarization that relies on computing the salience of each RDF sentence in terms of its centrality in the graph. In doing so, an RDF sentence graph is built and analyzed based on metrics such as degree centrality, shortest-path-based centrality and eigenvector centrality. A similar method for building personalized ontology summaries based on metrics coming from complex network theory is proposed in [36].

Other large semantic networks derived from the Web of Data have been examined as complex networks revealing their scale-free and small-world nature [20, 24] as well as their community structure [39, 13]. The vulnerability of the Web of Data from a complex network perspective and different mechanism for improving its robustness are discussed in [23]. "Folksonomies", different forms of social bookmarking, and other types of collaborative knowledge bases are special forms of semantic networks that can also be studied under the com-

[3] `http://www.daml.org/ontologies/`

plex network perspective, as discussed in [14, 47, 56, 25, 28].

The concept of Relevance is closely related to structural aspects of topical and semantic networks. In [22] relevance is defined as a measure of the information conveyed by a document regarding a query. Additionally, Rees & Saracevic [37] and more recently Saracevic [41] claim that relevance should take into account concepts such as the previous knowledge of the user and the usefulness of the information to the user. Keeping the user as the center of relevance measures, Barry [4] establishes additional criteria: information content of the document, the user's previous knowledge, the user's preferences, other information and sources within the environment, the document sources, the document as a physical entity, and the user's situation. Other aspects such as topicality, novelty, reliability, understandability, and scope are analyzed by Xu & Chen [54]. Hjorland [26] presents a review of relevance in the field of information science.

The concept of Relevance propagation has been applied for different purposes, such as identifying authorities in the Expert finding task [38]. Also, Topic distillation has benefited from relevance propagation, as shown by Chibane & Doan [15]. In these works, RPM's for hypertext document collections are computed in terms of content and link similarities, and the user's behavior, allowing to identify authoritative sources. Relevance propagation can be implemented by means of class evolution in topic ontologies. Such an evolution consists in adding new documents to the ontology classes. In [49] this idea is employed to guide focused crawlers. Kim & Candan [29] also apply this idea in a keyword propagation algorithm for augmenting the description of the entries in a navigation hierarchy.

4. Analysis

With the purpose of extending the analysis presented in [53], the measures described in the Background section were applied to multiple DMOZ RPM's. Some graphs corresponding to RPM's proposed in [52] are included, as well as new graphs derived from them.

4.1. Dataset

The principal data set comprises the basic DMOZ ontology. The DMOZ ontology used in our analysis comprises a set of 571,148 topics connected by 571,147 taxonomy links (T), 545,805 "symbolic" links (S) and 380,264 "related" links (R).

Table 1: Denomination, expression and number of edges of the DMOZ RPM's.

Denomination	Expression	Number of edges
G_1	$T \vee S \vee R \vee I$	2,068,364
G_2	$T \vee S \vee R \vee R^T \vee I$	2,269,866
G_3	T^*	4,588,580
G_4	$T^* \vee S \vee R$	5,502,581
G_5	S^*	3,817,557
G_6	$R \vee R^T \vee I$	1,301,060
G_7	$T \vee S \vee I$	1,688,100
G_8	$T^* \vee S$	5,122,366
G_9	$S \vee R \vee I$	1,497,217
G_{10}	$S \vee R \vee R^T \vee I$	1,700,060
G_{11}	$T^* \otimes (R \vee I)$	5,657,838
G_{12}	$T^* \otimes (R \vee R^T \vee I)$	6,547,256
G_{13}	$T^* \otimes S^*$	10,141,973
G_{14}	$T^* \otimes (S \vee R \vee I)$	7,072,930
G_{15}	$(S \vee R \vee I) \otimes T^*$	71,443,444
G_{16}	$T^* \otimes (S \vee R \vee I) \otimes T^*$	170,573,370
G_{17}	$T^* \otimes (S \vee R \vee R^T \vee I) \otimes T^*$	174,534,253
G_{18}	$[T^* \otimes (S \vee I) \otimes T^*] \vee [T^* \otimes (R \vee R^T \vee I)]$	14,177,359
G_{19}	$T^* \otimes (S \vee I) \otimes T^* \otimes (R \vee R^T \vee I)$	16,915,322
G_{20}	$T^* \vee S \vee R \vee R^T$	5,702,391

The RPM's in Table 1 are visually represented in Figure 3 using the graphical convention described in Figure 2. That figure denotes an additional graphical convention used to represent reflexive-transitive closures of some basic DMOZ ontology components: the reflexive-transitive closure of T, i.e. T^*, is represented by a triangle with solid lines, while for the reflexive-transitive closure of S (S^*), a triangle with dashed lines is used. The converse of the relation defined by the R component, i.e. the relations coming from the transpose of R (R^T), is also represented with a special convention consisting of red dash-dot curves.

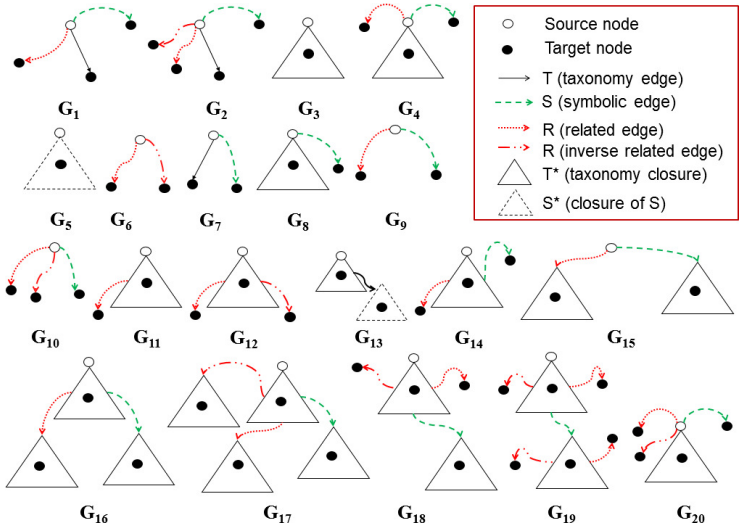

Figure 3: Graphical expression of the models of relevance propagation.

4.2. Connectivity and Centrality measures

The next subsections present the topological metrics used in this work to characterize the structure of the DMOZ graph and the RPM's. For some cases, such as Graph Density and Diameter, the complete set of measures are available as supplementary material.[4]

[4] `https://github.com/edus1984/Structural-Analysis-of-Relevance-Propagation-Models`

4.2.1. Graph Density and Diameter of RPM's

The graph density measure is reported in Figure 4. The values for T, S, R and G (G_1) are taken from [53]. The RPM's are sparse, with a density below 0.01% in most of the cases. The reflexive-transitive closure of T, namely T^*, results in a significantly increased density reflected by RPM's G_3, G_4, G_5 and G_8 with respect to prior models. Higher increases in density are observed when T^* is included with a Boolean product in a model, and even higher densities result from involving T^* on the right-hand side of the product. Models G_{14} and G_{15} represent the Boolean product of T^* and the union of the non-hierarchical components of DMOZ, where T^* is on the left- and right-hand side of the Boolean product, respectively. Note that the density of G_{15} is one order of magnitude higher than G_{14}, while G_{14} already shows an important increase in terms of density. The densest models are G_{16} and G_{17} because they involve T^* on both sides of the Boolean product, and the complete DMOZ graph in the middle, including the backward edges of R for the case of G_{17}. Models G_{18} and G_{19} do not exhibit very high densities because their Boolean products are performed between T^* and the basic models S and R, one at a time, but not simultaneously, as is the case with G_{16} and G_{17}.

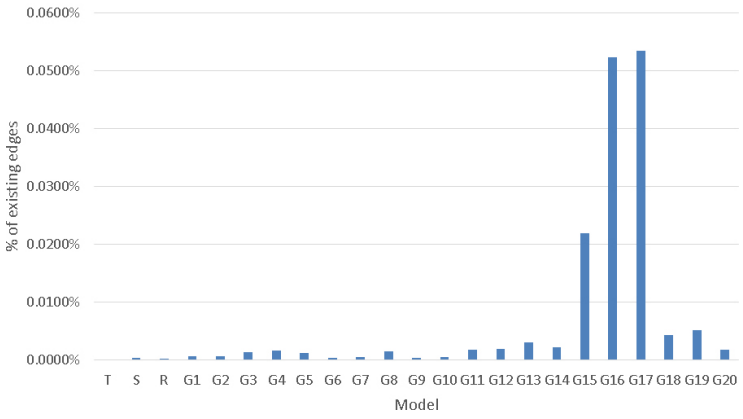

Figure 4: Graph densities of DMOZ RPM's.

Figure 5 shows the Diameter value for some representative RPM's. The

diameters of the basic DMOZ graphs (T, S, R and G_1) are described in [53]. The diameter of G_3 (T^*) is 1 given that all the possible paths in that model have a direct edge, as a result of the nature of a reflexive-transitive closure in a taxonomy. The same reasoning applies to G_5 (S^*). The largest diameters are associated with R and G_7, which are among the sparsest models. It is interesting to note how the diameter decreases when R is augmented with its backward edges resulting in model G_6. Such diameter decreases from 61 to 41, suggesting that the distance between some nodes can significantly reduce by only adding the backward edges of R. Note that the same behavior is also observed for G_1 and G_2. Apart from T, G_3 and G_5, models G_{19} and G_{13} have the lowest diameters. The RPM's with largest diameters are R, G_7, G_8 and G_9. Even though the high density of some RPM's can induce higher diameters, low diameter in sparser models can be a consequence of the existence of non-co-reachable nodes. Another noticeable phenomenon is the very high increase in diameter that results from augmenting G_3 (T^*) with S, resulting in G_8, and going from 1 to 58.

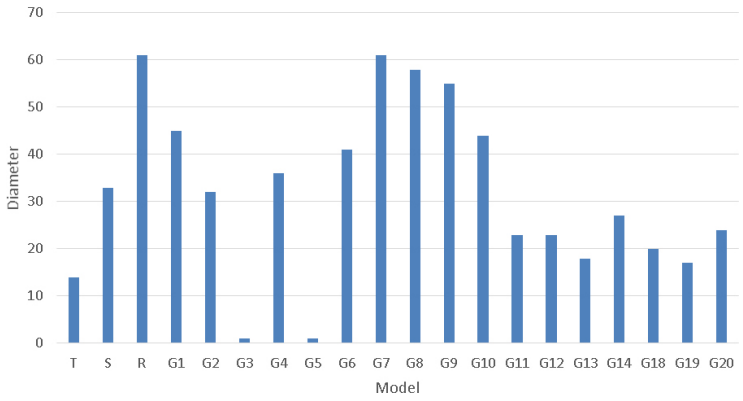

Figure 5: Diameters of DMOZ RPM's.

4.2.2. CPL, CL, Clustering Coefficient and Small-World Networks

Figure 6 shows the corresponding *CPL* and *CL* values for the subset of models with *CL* lower than 1,000, given that the differences in that measure are

not visible for greater values, and greater values are not relevant for a visual comparison. As can be seen in Table 2, models G_3, G_5 and G_{13} exhibit low CPL and very high CL values. This fact is a consequence of the existence of non-co-reachable nodes. It is also important to highlight the considerable decrease of CL when the R component is augmented with its backward edges (model G_6), falling in two orders of magnitude, from 4,997 to 76. The CPL and CL of G_7, G_8, G_9 and G_{10} are consistent, characterizing these RPM's as weakly connected networks. Finally, models G_{11}, G_{12}, G_{14}, G_{18}, G_{19} and G_{20} have low values of CPL and CL, i.e. suggesting that those networks are more connected than the rest. Particularly, the best connected model according to these measures is G_{19}, with a CPL of 4.86 and a CL of 9.

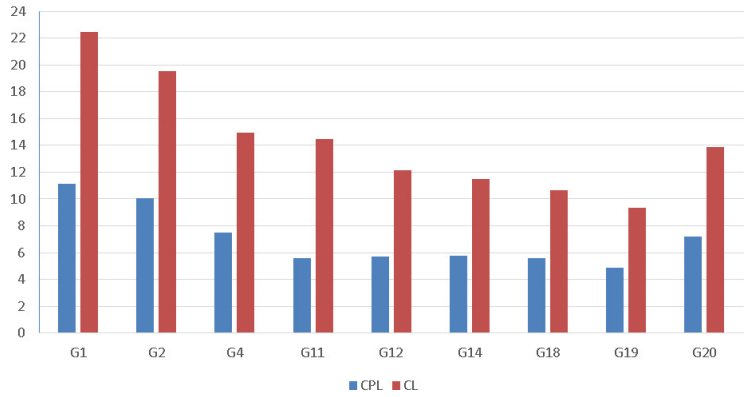

Figure 6: CPL and CL of the most representative DMOZ RPM's according to these measures.

Figure 7 shows charts with the grouped frequencies of CC values for each RPM. The CC values were grouped into 10 intervals of the same length between 0 and 1, for the purpose of showing how the values are distributed over each model. Some highlights about the CC values of the basic DMOZ graphs are provided in [53]. The highest average CC values are those corresponding to the RPM's that include the reflexive-transitive closure T^* of the taxonomy. In fact, the average CC is lower than 10% for the RPM's that do not include it, namely S, R, G_1, G_2, G_5, G_6, G_7, G_9 and G_{10}. Besides, the RPM's considered in this analysis exhibit 0 and 1 as CC while the nodes of the models including T^* never

Table 2: Graph density, diameter, *CPL* and *CL* of DMOZ models.

Model	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_1	0.0006%	45	11.1196	22
G_2	0.0007%	32	10.0808	20
G_3	0.0014%	1	0.8755	81,199
G_4	0.0017%	36	7.4666	15
G_5	0.0012%	1	0.8504	100,483
G_6	0.0004%	41	11.5456	76
G_7	0.0005%	61	19.1825	5,683
G_8	0.0016%	58	16.2893	5,683
G_9	0.0005%	55	14.3660	129
G_{10}	0.0005%	44	10.8486	48
G_{11}	0.0017%	23	5.5744	14
G_{12}	0.0020%	23	5.7070	12
G_{13}	0.0031%	18	4.2838	1,477
G_{14}	0.0022%	27	5.7619	12
G_{15}	0.0219%	-	-	-
G_{16}	0.0523%	-	-	-
G_{17}	0.0535%	-	-	-
G_{18}	0.0043%	20	5.5607	11
G_{19}	0.0052%	17	4.8597	9
G_{20}	0.0017%	24	7.1998	14

show *CC* values of 0 or 1. This phenomenon is related to the basic property of a reflexive-transitive closure of a taxonomy, that connects every node directly to its descendants. This process does not generate cycles, and turns non-immediate

descendants into direct neighbors. No edges between direct descendants are created, as seen in Figure 8. As a consequence, the CC of the parent node will be greater than 0 taking into account the edges that could arise between its descendants (children and grandchildren for instance), but will not be equal to 1, given that there are no edges traversing different branches of the taxonomy. This property holds for all the analyzed models that involve T^*.

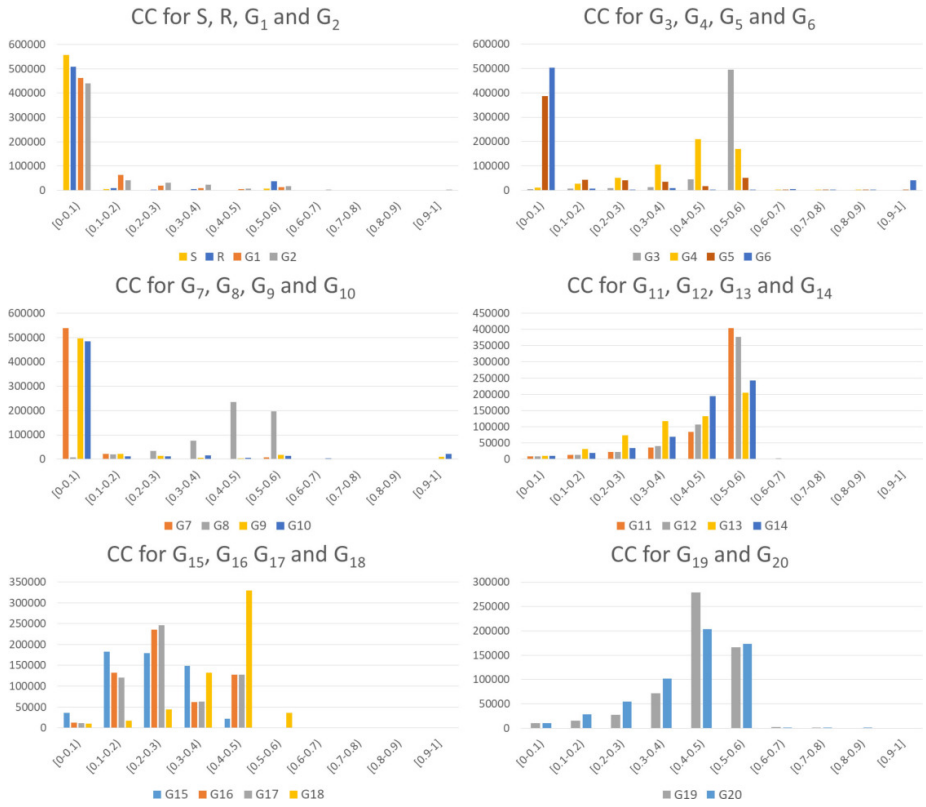

Figure 7: Grouped frequencies for Clustering coefficients of DMOZ RPM's.

Particularly, in models S, R, G_1, G_2, G_5, G_6, G_{10} most values agglomerate in the interval between 0 and 0.1. As it should be expected by its structure, the CC values of G_3 lie mostly in the 0.5-0.6 interval. For the case of G_{12}, G_{14}, G_{18} and G_{19}, the majority of CC values are in the interval 0.4-0.6. On the other hand, the values of G_4, G_8, G_{13}, G_{15}, G_{17} and G_{20} are more smoothly

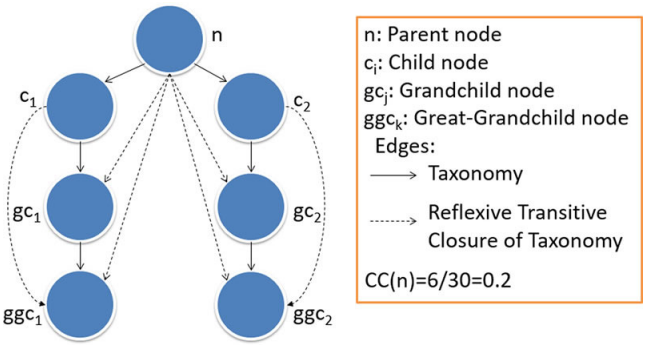

Figure 8: A regular taxonomy augmented with the edges of the corresponding reflexive-transitive closure, illustrating the CC value for the parent node n_i.

distributed. Another important issue to highlight is the very low number of nodes with CC values higher than 0.6 in all the RPM's.

As shown in previous studies [46, 8, 34], well-specified topic networks tend to group semantically related topics into cohesive communities and to directly or indirectly connect most of the topics through short meaningful paths. This behavior was already reported in our previous study of topic ontology networks [53] and is also observable in the RPM's analyzed here. More precisely, according to Figures 6 and 7, and considering the *Average Clustering Coefficient*, the CL and the CPL, the RPM's that better approximate a *small-world* network behavior are G_{19}, G_{18}, G_{12} and G_{14}, in that order. While the *small-world* phenomenon is expected to be present in any good RPM, short distances between topics may only represent a partial and indirect predictor of the model suitability. This is due to the absence of a necessary relation between the number of topological links in a path and the semantic distance between the corresponding topics.

4.3. Degree Distributions and Scale-Free structure

The charts in Figure 9 show in- and out-degree distributions in log-log scale for a set of representative RPM's. For instance, the in-degree distribution plots of the R component and the model G_1 -as seen in [53]- and the out-degrees of R, G_1 and G_3 indicate a possible power-law behavior. Following this notion,

the in-degree and out-degree distribution plots of G_{17} do not exhibit power-law distributions. Besides, it is important to note that, in contrast to the other plots in the figure, the in-degree distribution plot for G_3 is shown in a linear scale on both axes. That plot has the purpose of showing a possibly normal distribution of the in-degrees in G_3. As seen in the results of other computed network measures, this seemingly normal distribution for G_3 is the consequence of the structure of the reflexive-transitive closure of T.

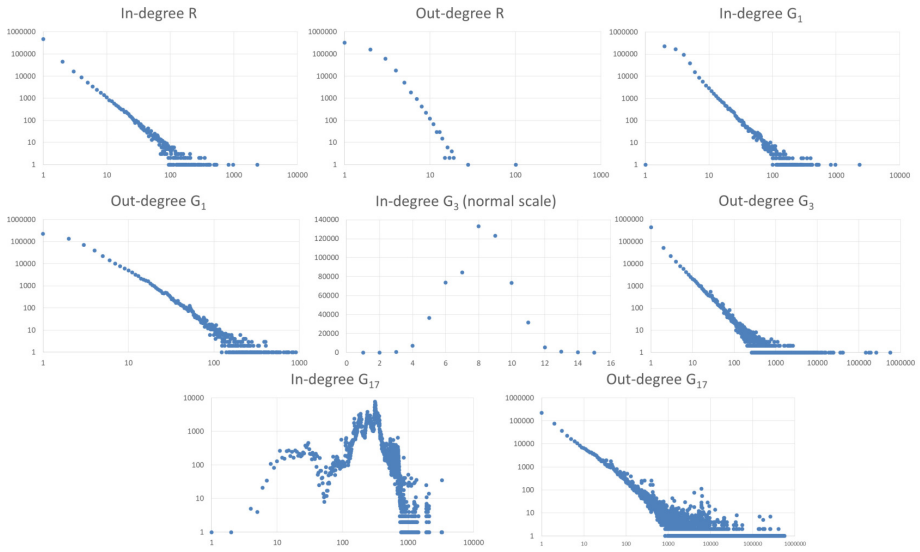

Figure 9: *In-degree* and *Out-degree* frequency distributions in log-log scale for R, G_1, G_3 and G_{17}.

The slight curvature exhibited by the out-degree plot of G_1 in Figure 9, as also explained in [53], could point out to an exponential rather than power-law behavior [16]. Such a curvature can be attributed to the taxonomy T, since the degree distributions for the RPM's that include that component exhibit such a shape. Also, the reflexive-transitive closure T^* causes linear increases in the in-degrees and exponential increases for the case of out-degrees in the RPM's that include it, stressing this effect on the distributions.

The scaling exponent of each distribution allows a more accurate analysis of the *scale-free* behavior of the networks. This exponent is illustrated in Fig-

ure 10. Most of the RPM's exhibit in-degree scaling exponents higher than 2, but the out-degrees have lower values in general. Particularly, models G_3, G_4 and G_5 and all the models beyond G_{10} exhibit lower values in their out-degree scaling exponents, and it could be due to the inclusion of T^* through Boolean products. The curvature described previously and the low scaling exponents in the out-degrees pose exponential distributions as better probabilistic models than power-law distributions. As stated in [53], the out-degree of R and the in-degree of S show the clearest power-law behaviors when the scaling exponents are considered, and the very high scaling exponent in the in-degree of R could be the effect of an arbitrary process of topic association, where some DMOZ collaborators added many "related" links to a few topics (e.g. the topic "Top/ Regional/ North America/ United States/ Arts & Entertainment/ Music" is linked from 2.341 nodes). Exponents associated with the most basic RPM's, namely S, R, G_1, G_2, G_6, G_7, G_9 and G_{10}, are higher than the exponents associated with the most sophisticated ones.

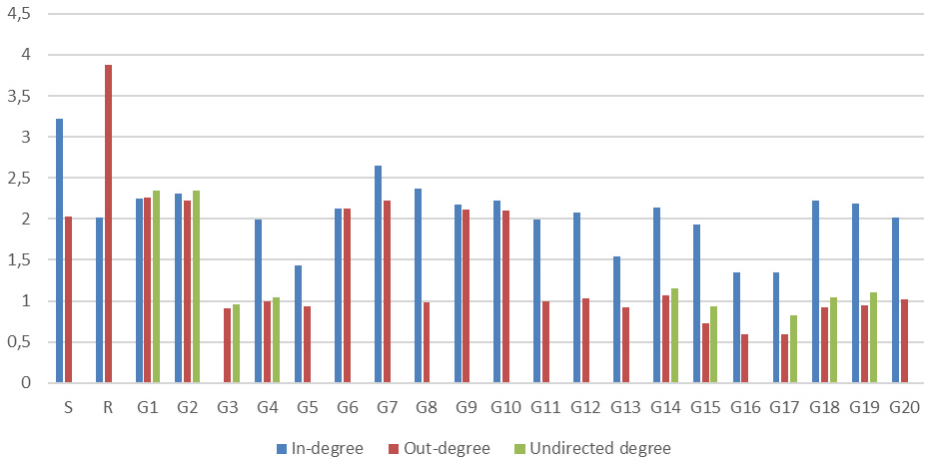

Figure 10: Scaling exponents of in-degree, out-degree and undirected graph degree distributions for DMOZ RPM's.

The edges in every RPM can be seen as non-directional relations between DMOZ topics, generating undirected graphs that can also be statistically analyzed. Figure 10 also shows the scaling exponents of the underlying undirected

graphs for some RPM's. Models G_1 and G_2 have scaling exponents greater than 2 and smaller than 3, indicating well-defined means and non-finite variances, as the most recognized power laws in nature [35]. Particularly, the unique cases of distributions with finite variance are the in-degree of S and the out-degree of R, considering their scaling exponents higher than three.

4.4. Visual analysis of centrality and connectivity

The visualizations of the RPM's were computationally very expensive to obtain, as undirected representations of the complete graphs were represented. To highlight some interesting facts, Figure 11 presents a visualization of G_1 (previously analyzed in [53]) and other RPM's using the *Large Networks Visualization* tool (LaNet-Vi [2]) to generate visualizations of the undirected graphs associated with the analyzed models employing a *k-core decomposition* algorithm [44]. A description of the criteria used to separate cores by the algorithm is available as supplementary material.[5]

In the visualization of G_1, separated cores and a low overlapping among them are observable (white gaps among the circles formed by the groups of nodes). Also, it is possible to distinguish small circles outside the border of the central core, representing smaller communities of topics. It is possible to observe differences in the overlapping of cores of G_3 and G_4. As model G_3 corresponds to the reflexive-transitive closure of T, there are more nodes with intermediate degrees, and the corresponding image shows a more stressed overlapping of cores. On the other hand, the participation of components S and R in G_4 induces a smoother structure of the cores.

The four images at the center and bottom of Figure 11 correspond to more complex RPM's, all of them including Boolean products of the reflexive-transitive closure of T and the basic components T and R. Gradually, from model G_{14} going through models G_{15} and G_{18} to G_{19}, a growth in the gap between the central core and the border circle of the image can be appreciated. Such a gap,

[5] `https://github.com/edus1984/Structural-Analysis-of-Relevance-Propagation-Models`

which will be called "central gap" from now on, can be explained as a result of the progressive Boolean products involving T^* applied to generate the mentioned RPM's. As seen for the density of the corresponding graphs, there is a considerable difference in the number of edges when T^* appears in a Boolean product on the left-hand side as in G_{14}, or on the right-hand side as in G_{15}, or on both hands, as in G_{18} and G_{19}. Particularly when T^* is present in the left-hand side of a Boolean product, edges are added linearly to the model on the right, but conversely the increase in the number of edges is exponential when T^* is on the right-hand side. This can be observed in the gap emerging in G_{15}, considering that the only difference between that model and G_{14} is the order of appearance of T^* in the corresponding Boolean products. The particular structure of the central gap in G_{15} reflects a set of peaks in the degree distribution of that model, since the graphic is much less dense than the one corresponding to G_{14}. It presents considerably more nodes in the central cores than G_{18} and G_{19}, and those nodes are distributed among more separated circles than in G_{14}. The images of models G_{18} and G_{19} denote a bigger central gap than G_{15}, and this is due to the presence of more nodes with higher degrees. Figure 12 provides further insights into these questions, showing the cumulative frequencies for the degree distributions of the four RPM's analyzed. The particular structure of the central gap in the k-core decomposition image of G_{15} corresponds to the associated plot in this figure, reflecting the absence of some subsets of intermediate degrees. Besides, the cumulative frequencies of G_{18} and G_{19} show a higher number of nodes with higher degrees, pointed by the slower growth of the curve in the plots.

4.5. Connectivity and Centrality of some relevant topics

The analyzed connectivity and centrality measures allow to recognize that some topics are central across diverse RPM's. This section identifies a set of topics that consistently exhibit high values of degree, betweenness centrality, closeness centrality, harmonic centrality and Lin's index in most of the generated RPM's. The Clustering Coefficient measure does not present significant

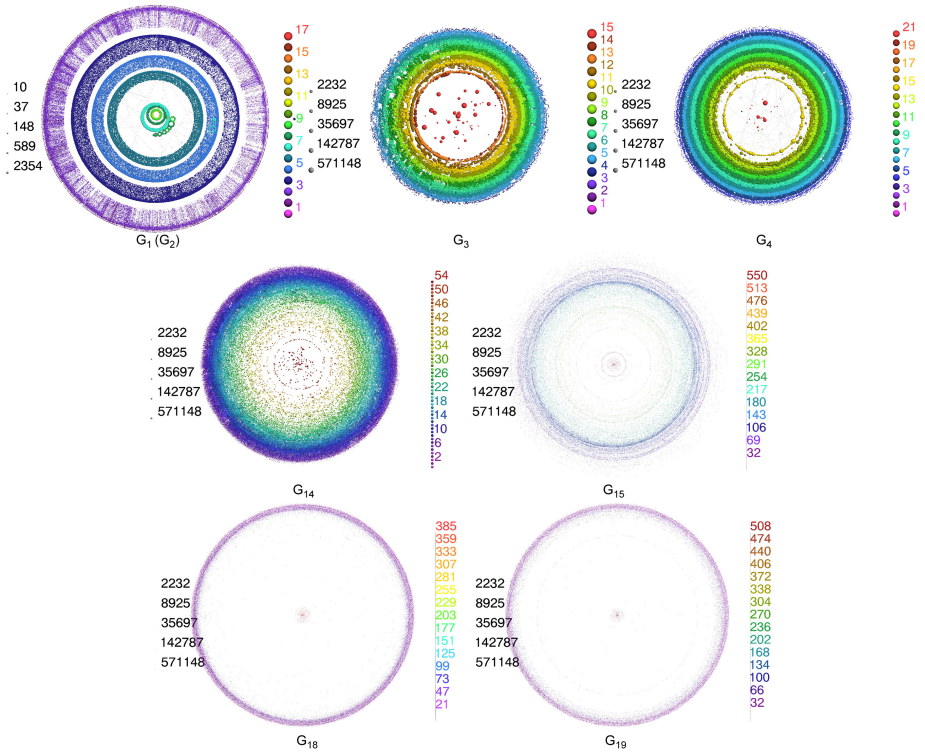

Figure 11: K-core decomposition visualizations of a selection of DMOZ models.

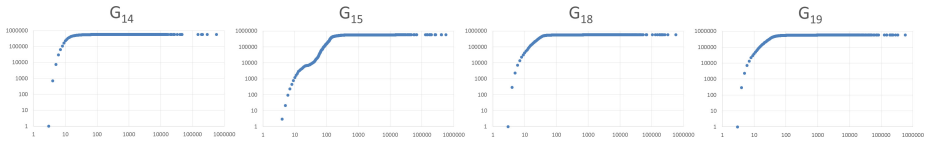

Figure 12: Cumulative Frequencies for degree distributions of G_{14}, G_{15}, G_{18} and G_{19} in log-log scale.

differences for any particular node in the RPM's, and therefore was not considered for this analysis. The results are summarized in Table 3.

As can be seen in Table 3 some topics that are central for the basic DMOZ graphs (e.g. "*Adult*") are also central for the derived RPM's. The high agreement in assigning topic "*Regional*" the highest value of BC in the majority of the studied RPM's lends this topic a relatively central position in the DMOZ ontology, when the shortest path length is considered as a basis for centrality. On the

Table 3: Most central topics of DMOZ RPM's. The rows correspond to the topics. The columns are associated with the connectivity and centrality measures considered. Each cell of the table enumerates the RPM's where the corresponding topic has the highest value.

Topic	Degree	Betweenness centrality	Closeness centrality	Harmonic centrality	Lin's index
Regional		$\mathbf{G_1}$,$\mathbf{G_4}$,$\mathbf{G_{11}}$, $\mathbf{G_{12}}$,$\mathbf{G_{14}}$, $\mathbf{G_{18}}$,$\mathbf{G_{20}}$		$\mathbf{G_6}$	$\mathbf{G_1}$,$\mathbf{G_6}$
Adult			$\mathbf{G_1}$,$\mathbf{G_2}$,$\mathbf{G_4}$, $\mathbf{G_{11}}$,$\mathbf{G_{12}}$, $\mathbf{G_{14}}$,$\mathbf{G_{18}}$, $\mathbf{G_{19}}$,$\mathbf{G_{20}}$		
Recreation/Travel		$\mathbf{G_6}$,$\mathbf{G_{10}}$		$\mathbf{G_2}$	$\mathbf{G_2}$
Science/ Environment		$\mathbf{R}$		$\mathbf{R}$	$\mathbf{R}$
Society/People/ Personal Homepages				$\mathbf{G_{11}}$,$\mathbf{G_{12}}$, $\mathbf{G_{14}}$,$\mathbf{G_{18}}$	$\mathbf{G_{11}}$,$\mathbf{G_{12}}$, $\mathbf{G_{14}}$,$\mathbf{G_{18}}$, $\mathbf{G_{19}}$
World/Español/ Artes/Cine	$\mathbf{S}$,$\mathbf{G_7}$, $\mathbf{G_8}$,$\mathbf{G_{13}}$			$\mathbf{S}$	$\mathbf{S}$
Regional/North America/United States/Arts & Entertainment/ Music	$\mathbf{R}$,$\mathbf{G_1}$, $\mathbf{G_2}$,$\mathbf{G_4}$ $\mathbf{G_6}$,$\mathbf{G_9}$, $\mathbf{G_{10}}$,$\mathbf{G_{11}}$, $\mathbf{G_{12}}$,$\mathbf{G_{14}}$, $\mathbf{G_{18}}$,$\mathbf{G_{19}}$, $\mathbf{G_{20}}$			$\mathbf{G_4}$,$\mathbf{G_{19}}$, $\mathbf{G_{20}}$	

other hand, when the alternative centrality measures involving co-reachability are taken into account, this topic remains central only for a few models and measures, while topics "*Adult*", "*Society/ People/ Personal Homepages*" and "*Regional/ North America/ United States/ Arts & Entertainment/ Music*" acquire a higher centrality for a considerable part of the RPM's. The case of topics "*Recreation/ Travel*" and "*Science/ Environment*" is particularly interesting. While the first seems to gain more centrality when the backward edges of component R are included, as in models G_2, G_6 and G_{10}, the second remains more central for the basic component R without its backward edges. Regarding topic "*Regional/ North America/ United States/ Arts & Entertainment/ Music*", apart from its high values of Harmonic centrality in some RPM's, it is pointed as the node with highest degree in many models.

5. Generalizability

The methodology presented in this work can be applied to any ontology but it becomes particularly interesting when applied to ontologies that are comprised of different types of relations. An example of such an ontology is the Gene Ontology (GO)[6]. GO provides structured knowledge about the functions of genes and gene products. It has been extensively used in the life sciences and has been continuously constructed and refined by a team of ontology developers that includes biologists and knowledge representation specialists. The GO knowledge base contains GO terms that refer to biological processes, molecular functions and cellular components. These terms are linked by relations of different kind: "is a", "part of" and "regulates".

The GO graph used in our analysis comprises a set of 47,199 GO terms connected by 71,305 "is a" taxonomy links (T_{is_a}) and 7,100 "part of" taxonomy links (T_{part_of}). While the GO graph is significantly smaller than the DMOZ topic ontology, it provides an interesting case for investigating the generalizabil-

[6]The Gene Ontology is hosted at `http://geneontology.org`.

ity of the proposed methodology.

In our analysis different operations were applied to the T_{is_a} and T_{part_of} components of the GO graph, resulting in the RPM's described in table 4.

Table 4: Denomination, expression and number of edges of the GO RPM's.

Denomination	Expression	Number of edges
GO_1	$T_{is\_a} \vee T_{part\_of}$	78,405
GO_2	$T^*_{is\_a}$	505,236
GO_3	$T^*_{part\_of}$	15,503
GO_4	$T^*_{is\_a} \vee T^*_{part\_of}$	520,706

Several topological metrics were computed to characterize the structure of some of the GO graph's components and the derived RPM's. These metrics are presented in table 5. As can be observed, all the representations are sparse with the highest density in models that include the T_{is_a} component or its corresponding reflexive-transitive closure $T^*_{is\_a}$. Note that the density of T_{is_a} is an order of magnitude higher than that of T_{part_of}. As can be seen in table 5 all models exhibit a relatively low *CPL* and *CL*. It is interesting to note that the difference in an order of magnitude between *CPL* and *CL* for T_{is_a} and GO_2 ($T^*_{is\_a}$) is a consequence of the existence of non-coreachable nodes. This results from the fact that the "is a" hierarchy of GO contains three sub-ontologies. GO_2 and GO_3 have diameters of size 1, as is also the case for the reflexive-transitive closure of the taxonomical component of the DMOZ graph. The other models exhibit diameters in the range 9-14. Models that incorporate additional components, such as those resulting from including the "regulates" relations, might significantly reduce the reported diameters. Regarding the *CC* metric, it is possible to observe that the highest values are achieved by those models that include the reflexive-transitive closure of the denser taxonomical component T_{is_a}, namely GO_2 and GO_4.

Figure 13 presents the visualization of the k-core decomposition of the basic

Table 5: Graph density, diameter, *CPL*, *CL* and *CC* of GO models.

Model	Density	Diameter	CPL	CL	CC
T_{is_a}	0.0032%	14	3.1898	0.3135	0
T_{part_of}	0.0003%	10	0.5153	1.9407	0
GO_1	0.0035%	14	3.3087	0.3022	0.0053
GO_2	0.0227%	1	0.9146	1.0934	0.3659
GO_3	0.0007%	1	0.2474	4.0423	0.0399
GO_4	0.0234%	9	1.1559	0.8652	0.3502

GO components T_{is_a} and T_{part_of} as well as the the k-core decomposition of GO_1 and GO_4 using LaNet-Vi. It is worth mentioning that for the sake of completeness we present the visualization of the k-core decomposition for T_{part_of}. However, due to the sparsity of the corresponding graph and the existence of a huge number of isolated components, the algorithm does not provide a representative visualization. On the other hand, the three separate components that correspond to the three sub-ontologies of GO can be visually appreciated in T_{is_a}, GO_1 and GO_4. It is interesting to appreciate that the three "is a" hierarchies predominate over the less dense "part of" hierarchy. Finally, we can observe that the degrees and shell-degrees of the nodes in the central layers are significantly higher in GO_4 as a result of applying the reflexive-transitive closure to the basic components of GO.

6. Conclusions and Future work

This article addresses the question of how topics relate to each other through relevance relations forming a complex semantic network. It presents new insights into the notion of relevance by moving from traditional perspectives to a complex network theory approach. Topological measures of connectivity and centrality were applied to the underlying graph of each model, namely Density, Diameter, *CPL* and Connectivity length. The nodes of each graph were also

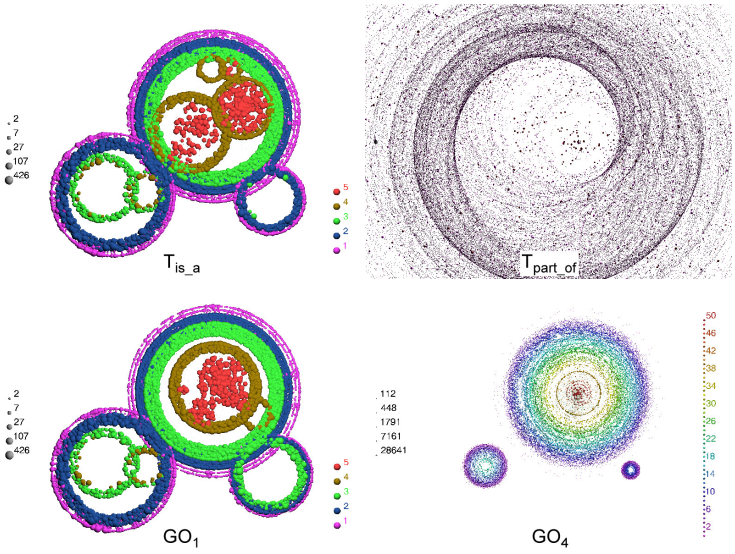

Figure 13: K-core visualizations of GO models.

the objective of local measures of centrality and connectivity, namely Clustering coefficient, Betweenness centrality, Closeness centrality, Harmonic centrality and Lin's index. Also, the degree distributions of each RPM were statistically and visually analyzed. According to the results obtained from these analyses, the underlying graphs of the RPM's display patterns commonly found in other complex networks. Similar to the conclusion reached in [53] for the DMOZ basic ontology, we can also state that the RPM's reveal small-world and scale-free features. While this work identifies the main topological characteristics of RPM's it did not look directly into the questions of what are the factors that produce a small-world topology or what are the mechanisms responsible for the emergence of scale-free features in these models. To answer the first question, we could assume that the small-world property results from the fact that all knowledge is related in some way, resulting in short relevance-relation paths between any pair of topics. Answering the second question from a network-growth perspective [3] requires modeling the network evolution through time by studying the dynamics of the knowledge engineering process that took place

in the construction of the topic ontology under analysis. Alternatively, to offer a simple feasible explanation for the observed scale-free topology we could conjecture that similar to many models of conceptual organization, the topology of the analyzed RPM's relies on the *cognitive economy principle* [40]. This principle underlies the hierarchical organization of conceptual knowledge where concepts with a higher number of semantic relationships are situated on the upper level of the knowledge representation structure [17]. Also, our analysis indicates that relations not necessarily coming from the hierarchical structure of topical ontologies play a key role. Hence, the high out-degree showed by some topics may, in part, be the result of topic generality (i.e., the topic is relevant to many descendant subtopics) but it may also be the consequence of other factors such as topic complexity. This is the case when a topic may be analyzed under multiple contexts or dimensions and therefore it becomes directly or indirectly related to several other topics. On the other hand, a high in-degree will be an indication of the pertinence of the topic to many other topics, either more general or related ones.

The analysis performed here is carried on the totality of the DMOZ network, overcoming the limitations that may result from sampling the dataset. The derived RPM's have equally weighted edges. An alternative approach would consist in using different weights for different type of links and to penalize these weights as paths become longer. The models can also be augmented with information derived from the content of the topics (e.g., terms and urls from the indexed webpages). This content can be used to adjust the weights of the links. The significance of centrality measures requires a prior characterization of the flow in a network [7]. In the scope of topic ontologies, it involves precise definitions of the way relevance and meaning are propagated through nodes in the graph. As a result of such analysis, weighted graphs could be derived. In this manner, a more appropriate "degree" of relevance could be acquired. This constitutes a proposal of future work to extend the present article.

Finally, we plan to derive networks of semantically connected topics and RPM's from other corpora, such as Wikipedia, and to apply an analysis similar

to the one introduced in this article. The increasing use of networked knowledge bases allows us to believe that this methodology will find many applications in the area of information science.

Acknowledgments

This work was supported by National Scientific and Technical Research Council (CONICET PUE 22920160100056CO), Ministerio de Ciencia y Tecnología (MinCyT, PICT 2016-0460, PICT 2019-03944), Universidad Nacional del Sur (PGI-UNS 24/N051) and Universidad Nacional de Salta (Proyecto CIUNSa A 2364 and Proyecto CIUNSa C 2659).

References

[1] F. N. Al-Aswadi, H. Y. Chan, K. H. Gan, Automatic ontology construction from text: a review from shallow to deep learning trend, Artificial Intelligence Review (2019) 1–28.

[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] C. L. Barry, User-defined relevance criteria: An exploratory study, Journal of the American Society for Information Science 45 (1994) 149–159.

[5] A. Bavelas, Communication patterns in task-oriented groups., Journal of the acoustical society of America 22(6), 725-730 (1950).

[6] N. D. Belnap, Entailment and relevance, The Journal of Symbolic Logic 25 (2) (1960) 144–146.

[7] S. P. Borgatti, Centrality and network flow, Social networks 27 (1) (2005) 55–71.

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

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

[10] 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.

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

[12] D. Buscaldi, D. Dessì, E. Motta, F. Osborne, D. Reforgiato Recupero, Mining scholarly data for fine-grained knowledge graph construction, in: CEUR Workshop Proceedings, vol. 2377, 2019, pp. 21–30.

[13] A. A. M. Caraballo, B. P. Nunes, G. R. Lopes, L. A. P. P. Leme, M. A. Casanova, Automatic creation and analysis of a linked data cloud diagram, in: International Conference on Web Information Systems Engineering, Springer, 2016, pp. 417–432.

[14] 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.

[15] I. Chibane, B.-L. Doan, Relevance propagation model for large hypertext document collections, in: Large Scale Semantic Access to Content (Text, Image, Video, and Sound), RIAO '07, Le centre de hautes etudes internationales d'informatique documentaire, Paris, France, 2007, pp. 585–595.

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

[17] A. M. Collins, M. R. Quillian, Retrieval time from semantic memory, Journal of Verbal Learning and Verbal Behavior 8 (2) (1969) 240 – 247.

URL http://www.sciencedirect.com/science/article/pii/S0022537169800691

[18] W. S. Cooper, A definition of relevance for information retrieval, Information storage and retrieval 7 (1) (1971) 19–37.

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

[20] W. Ge, J. Chen, W. Hu, Y. Qu, Object link structure in the semantic web, in: Extended Semantic Web Conference, Springer, 2010, pp. 257–271.

[21] R. Gil, R. García, J. Delgado, Measuring the semantic web, AIS SIGSEMIS Bulletin 1 (2) (2004) 69–72.

[22] W. Goffman, On relevance as a measure, Information Storage and Retrieval 2 (3) (1964) 201–203.

[23] C. Guéret, P. Groth, F. van Harmelen, S. Schlobach, Finding the achilles heel of the web of data: Using network analysis for link-recommendation, in: Proceedings of the 9th International Semantic Web Conference on The Semantic Web - Volume Part I, ISWC'10, Springer-Verlag, Berlin, Heidelberg, 2010, pp. 289–304.
URL http://dl.acm.org/citation.cfm?id=1940281.1940301

[24] C. Guéret, S. Wang, P. Groth, S. Schlobach, Multi-scale analysis of the web of data: A challenge to the complex system's community, Advances in Complex Systems 14 (04) (2011) 587–609.

[25] H. Halpin, V. Robu, H. Shepherd, The complex dynamics of collaborative tagging, in: Proceedings of the 16th International Conference on World Wide Web, WWW '07, ACM, New York, NY, USA, 2007, pp. 211–220.
URL http://doi.acm.org/10.1145/1242572.1242602

[26] B. Hjørland, The foundation of the concept of relevance, Journal of the American Society for Information Science and Technologycec 61 (2) (2010) 217–237.

[27] 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.

[28] J. Idrais, Y. El Moudene, A. Sabour, Online social networks: Study and validation of the regular behaviors of a targeted community on a complex network using a time series approach, in: 2019 4th World Conference on Complex Systems (WCCS), IEEE, 2019, pp. 1–7.

[29] J. W. Kim, K. S. Candan, Leveraging structural knowledge for hierarchically-informed keyword weight propagation in the web, in: International Workshop on Knowledge Discovery on the Web, Springer, 2006, pp. 72–91.

[30] J. Li, C. Zhang, H. Tan, C. Li, Complex networks of characters in fictional novels, in: 2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS), IEEE Computer Society, 2019, pp. 417–420.

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

[32] M. H. Mak, H. Twitchell, Evidence for preferential attachment: Words that are more well connected in semantic networks are better at acquiring new links in paired-associate learning, Psychonomic Bulletin & Review (2020) 1–11.

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

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

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

[36] P. O. Queiroz-Sousa, A. C. Salgado, C. E. Pires, A method for building personalized ontology summaries, Journal of Information and Data Management 4 (3) (2013) 236.

[37] A. Rees, T. Saracevic, The measurability of relevance, PB (Series), Center for Documentation & Communication Research, Western Reserve University, 1966.

[38] H. Rode, P. Serdyukov, D. Hiemstra, H. Zaragoza, Entity ranking on graphs: Studies on expert finding (2007).

[39] M. A. Rodriguez, A graph analysis of the linked data cloud, arXiv preprint arXiv:0903.0194 (2009).

[40] E. Rosch, Principles of categorization, Concepts: core readings 189 (1999).

[41] T. Saracevic, The notion of relevance in information science: Everybody knows what relevance is. but, what is it really?, Synthesis Lectures on Information Concepts, Retrieval, and Services 8 (3) (2016) 1–109.

[42] N. Sarshar, O. Boykin, V. Roychowdhury, Scalable percolation search on complex networks, Theoretical Computer Science 355 (1) (2006) 48–64.

[43] M. Savić, M. Ivanović, L. C. Jain, Analysis of ontology networks, in: Complex Networks in Software, Knowledge, and Social Systems, Springer, 2019, pp. 143–175.

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

[45] M. Shimono, N. Hatano, Efficient communication dynamics on macroconnectome, and the propagation speed, Scientific Reports 8 (1) (2018) 2510.

[46] 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.

[47] L. Stoilova, T. Holloway, B. Markines, A. G. Maguitman, F. Menczer, Givealink: mining a semantic network of bookmarks for web search and recommendation, in: Proceedings of the 3rd international workshop on Link discovery, ACM, 2005, pp. 66–73.

[48] S. H. Strogatz, Exploring complex networks, nature 410 (6825) (2001) 268.

[49] C. Su, Y. Gao, J. Yang, B. Luo, An efficient adaptive focused crawler based on ontology learning, in: Hybrid Intelligent Systems, 2005. HIS'05. Fifth International Conference on, IEEE, 2005, pp. 6–pp.

[50] Y. Theoharis, Y. Tzitzikas, D. Kotzinos, V. Christophides, On graph features of semantic web schemas, IEEE Transactions on Knowledge and Data Engineering 20 (5) (2008) 692–702.

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

[52] 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.

[53] E. Xamena, N. B. Brignole, A. G. Maguitman, A structural analysis of topic ontologies, Information Sciences 421 (2017) 15–29.

[54] Y. C. Xu, Z. Chen, Relevance judgment: What do information users consider beyond topicality?, Journal of the American Society for Information Science and Technology 57 (7) (2006) 961–973.

[55] S. Yoo, O. Jeong, Automating the expansion of a knowledge graph, Expert Systems with Applications 141 (2020) 112965.

[56] Y. Zeng, H. Wang, H. Hao, B. Xu, Statistical and structural analysis of web-based collaborative knowledge bases generated from wiki encyclopedia, in: Proceedings of the The 2012 IEEE/WIC/ACM International Joint

Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01, IEEE Computer Society, 2012, pp. 553–557.

[57] X. Zhang, G. Cheng, Y. Qu, Ontology summarization based on RDF sentence graph, in: Proceedings of the 16th international conference on World Wide Web, ACM, 2007, pp. 707–716.