Network Analysis of Semantic Web Ontologies

Conrad Roche
CS224W: Social and Information Network Analysis
December 11, 2011

Abstract

The semantic web community has introduced many, independently created, ontologies. These ontologies cover real-world domains, but are created and structured by humans.

This project aims to apply social network analysis to a graph representation of these ontologies. It aims to understand their structure and to see if they fit into any of the network models. It uses centrality to determine the important actors in the network and uses community detection techniques to understand the global structure of the ontology.

Background

Semantic Web

The semantic web aims to facilitate automated exchange of information in a machine consumable form [1]. Ontologies/Vocabularies define **concept**s and the relationships between them [2].

The Resource Description Language (RDF) specifies a way to state information in terms of statements [4] – which is a triple consisting of a "subject", "object" and a "predicate". The RDF Schema (RDFS) [5] allows us to define an RDF based vocabulary. The Web Ontology Language (OWL) is the ontology language for the semantic web [6].

The semantic web vocabularies are represented as graphs. In RDF, the predicate represents an edge between the subject and object nodes. RDFS also allows for relationship between predicates, preventing them from being viewed as pure edges from a network analysis perspective – since such relationships implies that there are edges between predicates.

Since the introduction of the semantic web technologies, many ontologies have been introduced. Many of these ontologies cater to a specific field or industry. Some of the ontologies, known as "upper ontologies" are more generic and cater to meta-information [7, 27]. Examples of upper ontologies include the Dublin Core (DC),

FOAF, SUMO and COSMO. Examples of domain-specific ontologies include OpenGALEN, and SWEET.

Prior work on network analysis of ontologies

Hoser et al. [8], performed social network analysis on the SUMO (Suggested Upper Merged Ontology) and SWRC (Semantic Web for Research Communities) ontologies. They found that social network analysis provide useful insights into the structure of ontologies. They found the need to preprocess ontologies to a simpler structure prior to the social network analysis. In this paper the authors explored the use of centrality analysis on the ontologies. They specifically identified betweenness centrality and eigenvector centrality for these two ontologies. The authors consider the betweenness centrality useful in identifying the core concepts in the ontology.

Stuckenschmidt [11] analyzed ontologies and used relative strengths to determine if an ontology needs to be partitioned. In the paper the author represented of the ontology as a proportional strength network where the weight of the relationship is determine by the inverse of the degree of the node. The partitions were then determined by applying minimal cut algorithm on the graph.

Coskun et al. [25] used social network analysis on ontologies to identify concept groups. In this paper the authors investigate nine (9) different representation of an ontology as a graph. The three basic representation being a plain RDF graph structure, a graph where the predicates are also represented as nodes and a third where only the classes are represented as nodes. Each of these representation had two extensions, one where the literals were ignored and another where the RDF, RDFS, OWL and XML Schema nodes were ignored.

Social network analysis has also been used for the development of ontologies [9], but that is not the focus of this project.

Approach taken for the Analysis of Ontologies

Ten ontologies were chosen for the analysis – of these 5 are RDFS based and the other 5 are OWL based. First the ontologies were pre-processed to transform them into a graph using Jena [28]. The graph was then analyzed using SNAP [29]. The identification of the nodes and edges for the graph is described in a later segment. The basic network properties determined for this graph are enumerated below.

The average degree was computed using $\bar{k} = \frac{2\,E}{N}$.
The network density was computed using E/N^2.

The diameter is the maximum (shortest path) distance between pair of nodes in the graph. This is computed using

$$\bar{h} = \frac{1}{E_{max}} \sum_{i,j \neq i} h_{ij}$$

Where h_{ij} is the shortest distance between nodes i and j.

The clustering coefficient is the fraction of a node's neighbors that are connected. The average clustering coefficient is computed using

$$C = \frac{1}{N} \sum_{i}^{N} C_i$$

$$= \frac{1}{N} \sum_{i}^{N} \frac{2\, e_i}{k_i\,(k_i - 1)}$$

Where e_i is the number of edges between node i's neighbors. k_i is the degree of node i.

Centrality

Degree centrality gives us a measure of how well connected a node is in a graph. A concept with many connections would be important and would exert a lot of influence on the graph. Any changes to concept (nodes) with high degree centrality will have a greater impact on end users of the ontology.

Betweenness centrality gives the normalized shortest path between nodes that pass through the given node. A large value would indicate that the node could reach other nodes in the network in fewer hops. In other words, these nodes behave as intermediaries for other nodes.

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Where, $C_B(v)$ is the betweenness centrality of node v; σ_{st} is the shortest path between nodes s, t; $\sigma_{st}(v)$ is the shortest path between nodes s, t that passes through node v; and V denotes the set of all the nodes in the graph.

Network Model

The scaling parameter (α) for a power-law network can be determined using the method of maximum likelihood (MLE). This is given by [26]

$$\hat{\alpha} = 1 + n \left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{min}} \right]^{-1}$$

Where $x_i, i = 1\ to\ n$ are the observed values of x (node degrees) such that $x_i \geq x_{min}$. Here x_{min} was taken as 1 for the computation in this analysis.

The standard error on $\hat{\alpha}$ is given by

$$\sigma = \frac{\hat{\alpha} - 1}{\sqrt{n}} + O(1/n)$$

Community detection

Perform community detection on the ontologies to understand its structure. If there are islands within the network – where the nodes are not connected to the rest – that would indicate that the ontology lacks cohesion. We can use WCC (weakly Connected Components) to help identify if the ontology contains any unrelated island of concepts.

We could also analyze the graph using the Clique Percolation Method (CPM) to detect closely related topics/concepts in the ontology.

Analysis of Ontologies

Ontologies analyzed

RDFS Schemas
This class of ontologies is described using the RDFS language. These tend to be smaller and more basic than the OWL based ontologies.

- FOAF (Friend of a Friend) is an ontology to describe details of a person. [14]
- WOT (Web of Trust) is an ontology to facilitate signing RDF documents. [15]
- DOAP (Description of a Project) is an ontology to describe software projects. [16]
- Dublin Core is an upper ontology from the Dublin Core Metadata Initiative. [13]
- AtomOwl is the ontology behind the Atom syndication format. [17]

OWL Ontologies

This class of ontologies is described using the OWL language. They are generally much more richer and use the more advanced concepts provided by OWL. The number of statements in OWL ontologies is usually an order of magnitude higher than those in RDFS Schemas.

- SWEET (Semantic Web for Earth and Environmental Terminology) is an ontology for environmental terms. [18]
- COSMO (Common Semantic Model) is a foundational ontology containing basic and primitive concepts. [19]
- OpenGALEN is an ontology to represent clinical information. The Common Reference Model (CRM) is the core of the ontology; the Diseases Extension is one of the sub ontologies. Version 8 of the ontology was used for this analysis. [20]
- SUMO (Suggested Upper Merged Ontology) is an upper ontology for general-purpose terms. [21]

The table below summarizes the number of statements, subjects, objects and properties in the above ontologies.

Ontology	Properties	Objects	Statements	Subjects
RDFS Schemas				
FOAF (Friend of a Friend)	82	263	1000	121
WOT (Web of Trust)	79	351	1174	147
DOAP (Description of a Project)	119	822	1881	178
Atom	169	607	2039	271
Dublin Core	85	567	2411	160
OWL Ontologies				
Semantic Web for Earth and Environmental Terminology (SWEET)	620	8743	78648	8494
COSMO	892	32159	380367	14669
Open GALEN CRM	1855	109790	595041	111804
Open GALEN - Diseases Extension	1901	284044	1359032	291579
SUMO	816	218226	2131244	91017

Identification of nodes and edges

The selection of nodes and edges to represent the ontology as a graph for analysis has been discussed before [22, 23, 25]. Some of the approaches taken are

a) Represent the plain RDF as a graph
Here the subject and the object are connected by a directed edge. The down side of this approach is that the predicates (which is a property) - that represent the edge - are ignored. While the same property could be present in the final graph due to statements about the property. This would result in an incomplete representation of the ontology.

b) Represent the predicates as nodes
Here the subject, object and predicates are represented as nodes. A statement/triple is converted into three edges – one from the subject to the predicate; second from the predicate to the object; and a third from the subject to the object.

Various variations are used on the above two graphs – with certain elements ignored. One variation is to ignore the literals; another is to consider only the classes and not the properties.

Another variation is to use the inferred model instead of the declared model for the analysis. The inferred model introduces new statements based on semantic reasoning on the ontology; while the declared model uses the statements explicitly declared in the ontology.

For the purpose of this analysis, option (b) was chosen – where the predicates are represented as nodes. Literals were ignored for this analysis – since they do not represent a concept or named entity in the model and are present for the purpose of description and documentation. The declared model was used for the purpose of this analysis. In other words, the nodes in the graph are the non-literal named resources in the ontology and the edges are the relationships between these resources as specified in the statements of the ontology.

Under the semantics of OWL, every class is a sub-class of itself. For the purpose of this analysis, these self-edges were ignored. Also, individuals (instance of class) were not considered for this analysis.

Basic Network properties of the ontologies

The table below summarizes some of the computed basic properties of the graph representation of the selected ontologies.

Ontology	#Nodes	#Edges	Avg. Degree	Density	Avg. Clustering Coeff.
RDFS Schemas					
FOAF (Friend of a Friend)	127	1036	16.31	6.42E-02	0.615
WOT (Web of Trust)	149	1234	16.56	5.56E-02	0.781
DOAP (Description of a Project)	182	1505	16.54	4.54E-02	0.644
Dublin Core	310	2183	14.08	2.27E-02	0.811
Atom	275	1955	14.22	2.59E-02	0.622
OWL Ontologies					
Semantic Web for Earth and Environmental Terminology (SWEET)	7312	68471	18.73	1.28E-03	0.781
COSMO	9872	291267	59.01	2.99E-03	0.669
Open GALEN CRM	30022	267909	17.85	2.97E-04	0.677
Open GALEN - Diseases Extension	40795	327129	16.04	1.97E-04	0.672
SUMO	258184	2587038	20.04	3.88E-05	0.824

Most of the ontologies, with the exception of COSMO, have an average degree between 13 and 20. These networks are sparse – but less sparse than other real world network [24], which have the average degree as less than 10. The network density is also much higher than other real world networks. The average diameter for all the ontologies was less than 3.5.

Ontology	Avg. diameter	Effective diameter	Max diameter
RDFS Schemas			
FOAF (Friend of a Friend)	1.47	1.98	5
WOT (Web of Trust)	1.63	2.54	4
DOAP (Description of a Project)	1.58	2.35	5
Dublin Core	1.85	2.37	5
Atom	1.46	2.04	5
OWL Ontologies			
Semantic Web for Earth and Environmental Terminology (SWEET)	2.16	3.59	16
COSMO	2.36	3.57	11
Open GALEN CRM	1.27	1.81	9
Open GALEN - Diseases Extension	1.31	1.87	9
SUMO	3.50	5.80	30

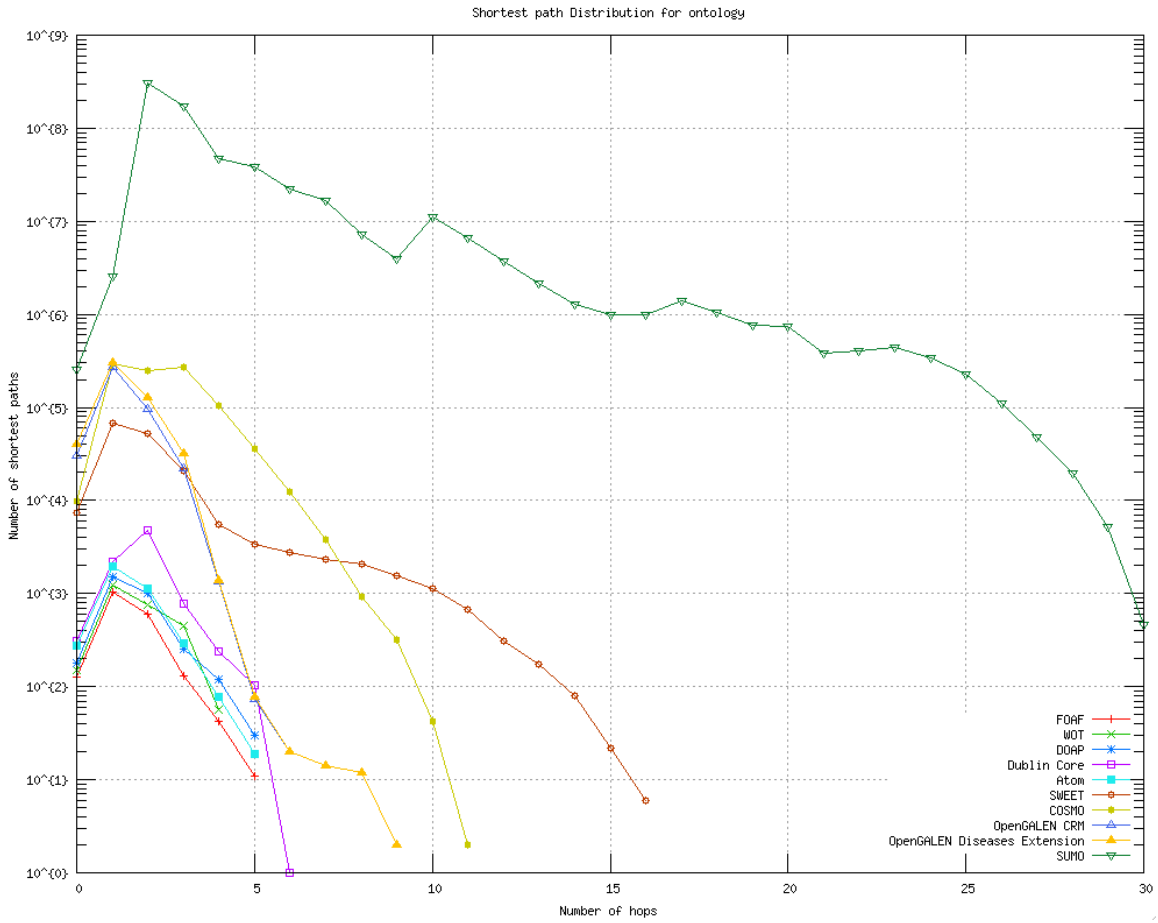

Fig 1. Shortest path distribution for the ontologies.

The above observation can be explained by the fact that ontologies focus on a specific field of study or topic and hence have a lot more relations between the nodes than a real world network will have.

The low diameter of these networks is due to the presence of supernodes like rdf:type and rdfs:subClassOf which have very high degrees. For instance, SUMO had 1,587,545 statements (76% of total) with the predicate as rdf:type; SWEET had 36,676 statements (48%) with the predicate as rdf:type and 31,441 (41%) with rdfs:subClassOf. Due to such high degrees they reduce distance between the nodes – hence reducing the diameter.

Network Model

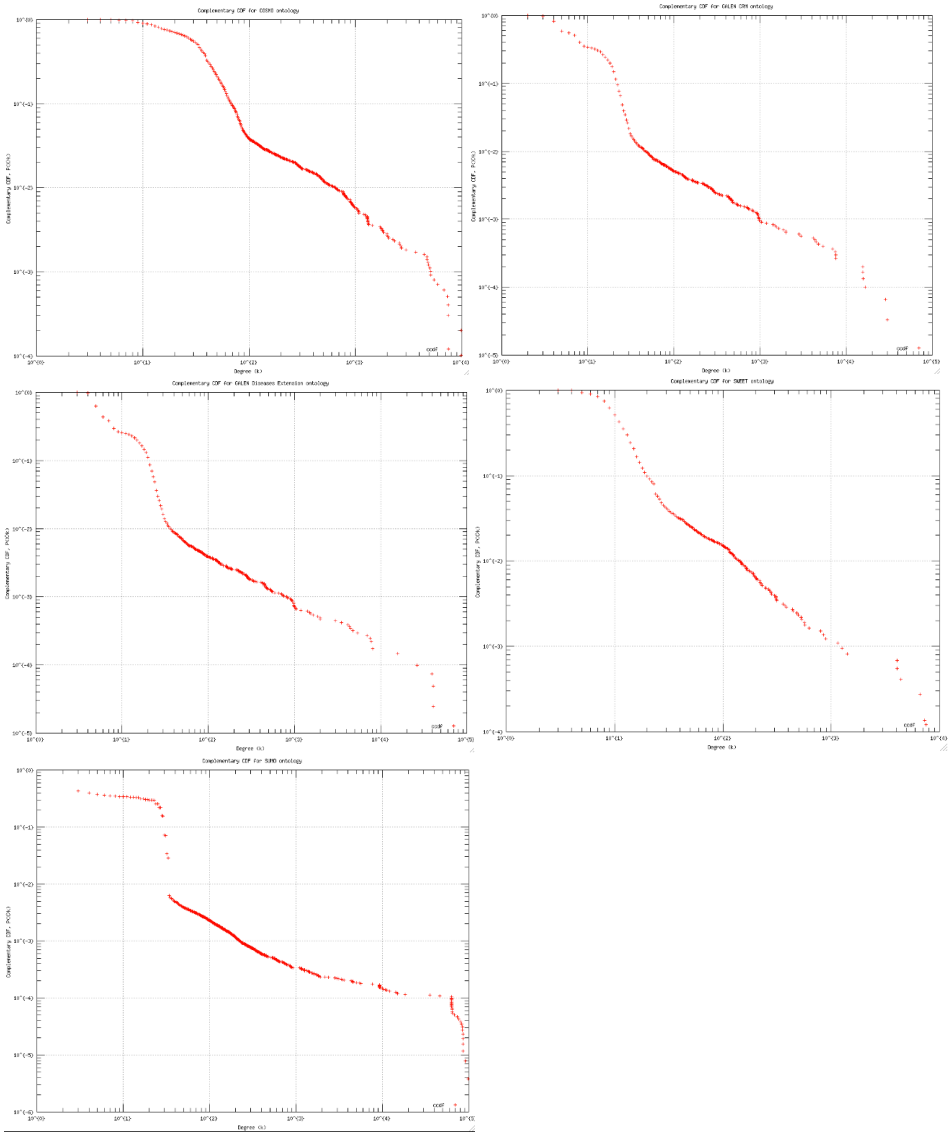

Fig 2. Complementary CDF Distributions (on a log-log scale) for the COSMO, OpenGALEN CRM, OpenGALEN Disease Extension, SWEET and SUMO ontologies.

The CCDF distributions for the ontologies show a prominent linear plot; indicating that these graphs follow the power-law network model. Using MLE these graphs were determined to have a scaling parameter around 1.8.

Ontology	Scaling factor (α)	Standard error (σ)
Semantic Web for Earth and Environmental Terminology (SWEET)	1.88	0.066
COSMO	1.82	0.042
Open GALEN CRM	1.80	0.053
Open GALEN - Diseases Extension	1.79	0.052
SUMO	1.86	0.041

Centrality

The betweenness centrality correctly identified some of the core concepts in the ontologies. For instance it identified Agent and maker as one of the core concepts for FOAF; Individual and Context for COSMO; relation, creator for Dublin Core; Substance, Quantity for SWEET; TopCategory, DomainCategory for OpenGalen CRM.

Due to the use of many of the central concepts from the OWL language, many of the OWL constructs show up as the central concepts for may of the ontologies. Some example of such nodes are - rdf:type, rdfs:Resource, rdf:Property, rdfs:isDefinedBy, rdfs:seeAlso, rdfs:domain and rdfs:range. To prevent this from interfering in the analysis, an approach can be to filter the results with the namespace of the ontology. For instance, filtering the result for DOAP correctly identifies the core concepts as Project and Repository. Without the filter, none of the concepts from DOAP show up on the top ten values.

The degree centrality analysis across the ontologies also gives a good insight into the usage of RDF, RDFS and OWL constructs. Based on the analysis of the 10 ontologies, the most widely used constructs are rdf:type, rdfs:Resource, owl:Thing, rdfs:seeAlso, and rdfs:isDefinedBy.

Community detection

All of the ontologies had only one WCC (weakly connected components) each, so there is no isolated island of concepts in these ontologies.

The Clique Percolation Method (CPM) was very effective in identifying communities within the ontology graphs. It correctly grouped together all the elements and compounds in the SWEET ontology; all the plexus (part of nervous system) in the OpenGALEN CRM ontology.

Conclusion

There are multiple ways in which the ontology can be represented as a graph. For the purpose of this analysis, the predicate was represented as a node in the graph representation. Network analysis of the ontologies generated useful results. Since the ontologies have specific focus areas, they have higher average degrees. Presence of supernodes reduces the average diameter of these ontologies to below 3. The basic network properties are useful in comparing ontologies in terms of complexity and level of detail.

Betweenness centrality was useful in identifying the central concept in the ontologies. It was also used to identify the core constructs in RDF, RDFS and OWL.

It is interesting to note that, even though the ontologies were developed independently, they follow the power-law and their scaling factor are similar (around 1.8). In the future, we can investigate to see if this hold for other semantic web ontologies.

The Clique Percolation Method (CPM) was very effective for grouping together closely related concepts.

Future analysis could consider different representation of the graph for the ontology and identify the best representation for the analysis. We could also investigate using the declared vs. the inferred model for the analysis. We could also analyze the strength of relationship between the nodes – as described by Stuckenschmidt [11] – to perform community detection.

References

[1] Tim Berners-Lee. Semantic Web Road map. 1998; http://www.w3.org/DesignIssues/Semantic.html

[2] W3C. Vocabularies. http://www.w3.org/standards/semanticweb/ontology, Accessed: Oct 26, 2011.

[3] T. Berners-Lee, R. Fielding, and L. Masinter. RFC 2396. Uniform Resource Identifier (URI): Generic Syntax. 1998.

[4] F. Manola and E. Miller. RDF Primer. 2004. http://www.w3.org/TR/2004/REC-rdf-primer-20040210/

[5] D. Brickley and R. V. Guha. RDF Vocabulary Description Language 1.0: RDF Schema. 2004. http://www.w3.org/TR/2004/REC-rdf-schema-20040210/

[6] P. Hitzler, M. Krötzsch, B. Parsia, P. F. Patel-Schneider, and S. Rudolph. OWL 2 Web Ontology Language Primer. 2009. http://www.w3.org/TR/2009/REC-owl2-primer-20091027/

[7] James Schoening. IEEE P1600.1 Standard Upper Ontology Working Group (SUO WG) Home Page. http://suo.ieee.org/, Accessed: Oct 26, 2011.

[8] Hoser, B et al. "Semantic Network Analysis of Ontologies." Proceedings of the 3rd European Semantic Web Conference 4011 (2006) : 514-529.

[9] Peter Mika. Ontologies Are Us: A Unified Model of Social Networks and Semantics. In Yolanda Gil, Enrico Motta, V. Richard Benjamins, and Mark A. Musen, editors, ISWC 2005, volume 3729 of LNCS, pages 522–536, Berlin Heidelberg, November 2005. Springer-Verlag.

[10] Newman, M. E. J.: Fast algorithm for detecting community in networks. Phys. Rev. E 69, 066133 (2004)

[11] Heiner Stuckenschmidt. Network Analysis as a Basis for Partitioning Class Hierarchies. In Proceedings of the ISWC2005 Workshop on Semantic Network Analysis (2005).

[12] Newman, M E J. "Finding community structure in networks using the eigenvectors of matrices." Physical Review E - Statistical, Nonlinear and Soft Matter Physics 74.3 Pt 2 (2006).

[13] DCMI Usage Board. DCMI Metdata terms. 2010; http://dublincore.org/documents/dcmi-terms/

[14] Brickley, D. & Miller, L., 2010. FOAF Vocabulary Specification. Namespace Document, 3(Revision 1.113), p.http://xmlns.com/foaf/spec/.

[15] Web of Trust RDF Ontology. http://xmlns.com/wot/0.1/

[16] Description of Project (DOAP). 2008; http://trac.usefulinc.com/doap

[17] AtomOwl. 2006; http://bblfish.net/work/atom-owl/

[18] Semantic Web for Earth and Environmental Terminology (SWEET). 2011; http://sweet.jpl.nasa.gov/ontology/

[19] Common Semantic Model (COSMO). 2009; http://ontolog.cim3.net/cgi-bin/wiki.pl?COSMO

[20] Rector, A.L. et al., 2003. OpenGALEN: open source medical terminology and tools. AMIA Annual Symposium proceedings AMIA Symposium AMIA Symposium, 2003, p.982.

[21] Niles, I. & Pease, A., 2001. Towards a standard upper ontology C. Welty & B. Smith, eds. Proceedings of the international conference on Formal Ontology in Information Systems FOIS 01, 2001, p.2-9.

[22] Stuckenschmidt, H. & Klein, M., 2004. Structure-based partitioning of large concept hierarchies. In In Proc 3rd International Semantic Web Conference ISWC2004. Springer, pp. 289-303.

[23] Schmitz, C., 2007. Self-Organized Collaborative Knowledge Management. , p.193. Available at: http://www.upress.uni-kassel.de/online/frei/978-3-89958-325-0.volltext.frei.pdf.

[24] J. Leskovec, K. Lang, A. Dasgupta, M. Mahoney. Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters. Internet Mathematics 6(1) 29--123, 2009.

[25] G. Coskum, M. Rothe, K. Teymourian, A. Paschke. Applying Community Detection Algorithms on Ontologies for Identifying Concept Groups. In O. Kutz and T. Schneider (Eds.) Modular Ontologies, IOS Press, 2011, pp. 12-24.

[26] A. Clauset, C.R. Shalizi, and M.E.J. Newman, "Power-law distributions in empirical data" SIAM Review 51(4), 661-703 (2009).

[27] Mascardi, V., Cordì, V. & Rosso, P., 2007. A comparison of upper ontologies. In M. B. E. Al, ed. Group. Citeseer, p. 55–64.

[28] HP Labs. Jena – A Semantic Web Framework for Java. http://openjena.org.

[29] Stanford. SNAP network analysis library. http://snap.stanford.edu