MEASURING THE SEMANTIC WEB

ROSA GIL
Universitat de Lleida, Jaume II 69, E-25001 Lleida, Spain
rgil@diei.udl.es

ROBERTO GARCÍA
Universitat Pompeu Fabra, Circumval·lació 8, E-08003 Barcelona, Spain
roberto@rhizomik.net

The goal of this work is double. First, to provide interoperability between two scientific disciplines that were not related until now, Complex Systems and Semantic Web. Second, to provide a model for Semantic Web behaviour based on complex systems properties. It is difficult to build such a model with "classical" tools. This is because the Semantic Web is a knowledge system that, unlike previous ones, is open and growths freely, without central control. The World Wide Web is based on the same principles. Both systems, the WWW and the Semantic Web, are acquiring a size and a complexity that puts them out of our control and even from our direct conception. It has been already shown that the WWW is a Complex System. We use graphs and statistical mechanics to model the Semantic Web behaviour. Our results are based on a pair of studies of Web ontologies collected from the Semantic Web performed during the last two years. The properties of the graph models we have built show that the Semantic Web is also a Complex System as it shares the same statistical patterns with other Complex Systems.

1 Introduction

The book *Weaving the Web* by Tim Berners-Lee [1] presents a plan to build what is called The Web. Basically, it can be described as a decentralised knowledge system that self-organises and evolves scaling to unforeseen conditions. The Semantic Web effort is introduced as the last step towards completing this idea, from the result of the World Wide Web effort.

This knowledge system is not like previous ones. It is open and growths freely, without central control, and this can produce many undesired outcomes that can be also seen as opportunities. The World Wide Web is based on the same principles, e.g. there are link consistency problems, but has largely succeeded.

However, these are only vague words. What have we really built with the World Wide Web? And what are we building with the Semantic Web? How near we are from the original plans and what is the metric? These are difficult questions. The WWW and the Semantic Web are acquiring a size and a complexity that puts them out of our control and even from our direct conception.

What can we do? Just looking around we realise that the WWW and the Semantic Web are just as complex as many other systems. Other research

communities have faced similar problems and found a common approach, which has properly been called the study of complex systems. Are they complex systems?

Complex systems are made up of the combination of a great amount of elements. However, their behaviour is not the sum of the behaviour of their parts. Examples of Complex Systems are metabolic networks [2], acquaintance networks [3], food webs [4] or neural networks [5]. Actually, it has also been shown that the WWW is a Complex System [6].

What do all these systems have in common? How to identify a Complex System? Scientists have looked for a way to achieve this using some mathematical tools, concretely graphs and statistical mechanics. In the next section this methods are presented.

2 Modelling and Analysing Complex Systems

Graphs are used to model Complex Systems in order to analyse them. Nodes represent the Complex System parts (chemical components, people, species, neurons, web pages…). Edges model the relationships among the parts (chemical reactions, acquaintanceship, species dependences, neuron axons, web links…).

The resulting graphs show statistical properties that characterise Complex Systems. Some of them are highlighted here. They are considered sufficient conditions for identifying a Complex System:

- **Degree distribution**: the resulting graphs, although they model systems that are shaped without a central control, are not random graphs, as it was first believed. The probability $P(k)$ that a vertex has a degree k does not follow a Poisson distribution as in random graphs.
  Instead, it shows a power-law distribution, $P(k) \approx k^{-r}$. This kind of distributions are characterised by the γ exponent and are called scale-free networks [7]. In other words, they show the same properties independently of the scale at which they are observed.
- **Small world**: a graph is a small world if the average minimum path length d between vertices is short [8,9], usually scaling logarithmically with the total number of vertices. Graphs showing an average path length similar to random graphs of the same size and average degree are very likely small worlds [10], $d \approx d_{random}$.
- **Clustering coefficient**: It measures the probability that two neighbours of a given node are also neighbours of one another. For random graphs it is a small quantity. However, Complex Systems show a high clustering compared to random graphs, $C >> C_{random}$. A high clustering confirms small-worldness.

3 Is the Semantic Web a Complex System?

We are now going to study the Semantic Web as a Complex System. It is modelled as a graph and then analysed using the statistical methods already presented. The results are analysed in order to check if it is a Complex System and to compare it with other ones. All the tools that have been used and the complete results are available at the project web page[a].

3.1 The Semantic Web Graph

The first step towards analysing the Semantic Web is to build an appropriate graph model. Due to self-similarity and scale invariance of Complex Systems, we can perform this analysis selecting a significant portion of the Semantic Web and the results can be inferred to other scales.

We have focused on the ontological part of the Semantic Web, i.e. we model the graph from a set of semantic web ontologies. We could also use instance metadata but we consider that, at this first stage, to focus on ontologies makes the conclusions more relevant.

Instance metadata usually models "real networks" that should be analysed on their own or have already been shown to be Complex Systems. For instance, FOAF metadata models social networks that have been extensively studied as Complex Systems [8].

Therefore, in order to collect the semantic web ontologies that are analysed, we have modified an existing RDF crawler[b] in order to facilitate the collection of a great amount of RDF metadata. This crawler has been launched over the DAML Ontology Library[c]. The processed URIs are combined in a RDF graph built in the last 2005 study from 1,365,286 triples for 282 ontologies at the DAML Ontology Library.

---

[a] Living Semantic Web project web page, http://dmag.upf.es/livingsw

[b] RDF Crawler, http://dmag.upf.edu/livingsw/nrdfcrawler.htm

[c] DAML Ontology Library web page, http://www.daml.org/ontologies

3.2 Graph analysis

In order to analyse the obtained Semantic Web graph we use Pajek[d], a large networks analysis tool. The RDF triples are translated to Pajek network format. The triples subjects and objects became network nodes connected by directed edges from subject to object.

For this first analysis we will focus on the explicit nature of the Semantic Web. Only triples explicitly stated in the processed Ontologies are considered. Therefore, for the moment, the potential triples that could be inferred applying RDF, DAML+OIL or OWL semantics are ignored.

The original Pajek network had 56,592 nodes and 131,130 arcs. However, in the last study in 2005, the Pajek network from DAML Ontology Library has 307,231 nodes and 588,890 arcs. Once loaded in Pajek, the available tools are used to obtain the required information about the graph:

- **Average degree and degree distribution**: use the Net/Partitions/Degree command. Then, generate the Vector for the Degree Partition and from Info/Vector get the mean value.
- **Clustering factor**: use the Net/Vector/ClusteringCoefficients/CC1 command to compute the 1-neighvourhood clustering coefficient for a directed graph. Then multiply the resulting mean value by two in order to compute it for an undirected graph.
- **Average minimum path length**: average over a random selection of 20 nodes (using Partition/CreateRandomPartitions and Partition/MakeCluster of size 20) and the averages of their k-neighbours vectors (using the Net/k-Neighbours with the Net/k-Neighbours/FromCluster option).
- **Power-law tails exponent**: linear regression from the degree distribution using GNUPlot[e].

All the results of the graph analysis are summarised in the next section.

3.3 Results

The results of the graph analysis are shown in Table 1. The first line, DAMLOntos, shows the results for the graph built from the ontologies at DAML library. It can be compared with the same parameters for other Complex Systems networks: the results from some WWW studies [11,6], WordNet [12] and human language words networks [13].

---

[d] Pajek, http://vlado.fmf.uni-lj.si/pub/networks/pajek

[e] Gnuplot Central, http://www.gnuplot.info

Table 1. Some Complex System statistical properties. Networks, number of nodes, average degree <*k*>, clustering factor *C*, average path length <*d*> and power-law exponents γ

Network	Nodes	<k>	C	<d>	γ
DAMLOntos (2003-4-11)	56,592	4.63	0.152	4.37	-1.48
DAMLOntos (2005-1-31)	**307,231**	**3.83**	**0.092**	**5.07**	**-1.19**
WWW	~200 M		0.108	3.10	-2.24
WordNet	66,025		0.060	7.40	-2.35
WordsNetwork	500,000		0.687	2.63	-1.50

First of all, from the previous data, we can deduce that the Semantic Web is a small world comparing its average path legth <d>=4.37 to the corresponding value for a random graph with the same size and average degree, $<d>_{rand}$=7.23. Moreover, the clustering factor C=0.152 is much greater than C_{rand}=0.0000895 for the corresponding random graph.

The final evidence is the degree distribution; it is clearly a power-law. The degree Cumulative Distribution Function (CDF) for the older DAMLOntos has linear regression with an exponent γ = -1.485 with a regression error $\varepsilon_{\%}$ = 1.455. In the last study of DAMLOntos, the linear regression of this function gives an exponent γ = -1.186 with a regression error ε% = 0.896, Fig. 1.

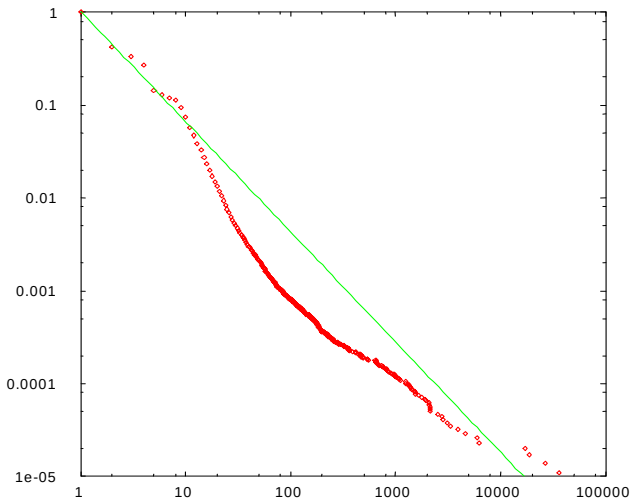

Fig. 1. Degree CDF (Cumulative Distribution Function) for the last study of the DAML library ontologies (DAMLOntos) plus linear regression and computed exponent

Therefore, the graph for the portion of the Semantic Web that has been analysed shows clear evidences that the Semantic Web behaves like a Complex System. It is a small world, with a high clustering factor and a power-law degree distribution. It has also a scale-free nature, so the same properties can be observed at a different scale.

Indeed, the analysis has been repeated for smaller graphs yielding the same conclusion. For instance, for a 971 nodes graph corresponding to the IPROnto [14] ontology: $C = 0.071$ while $C_{rand} = 0.0034272$, $\langle d \rangle = 3.99$ while $\langle d \rangle_{rand} = 5.38$ and $\gamma = -1.06$ with $\varepsilon_{\%} = 4.45$.

4 Conclusions and future work

It has been shown that the Semantic Web behaves like a Complex System. When it is viewed as a graph, it reproduces all the characteristic patterns that all Complex System share. Once the Semantic Web is studied from this perspective, these patterns can be used as a kind of Semantic Web metric. With them, we can figure out its current situation and compare it to other Complex Systems.

We have just started this work and a lot of questions have emerged. We plan to apply inferences to the retrieved triples in order to check the resulting graph. What do the implicit semantics do from the perspective of the whole RDF graph? Instance metadata is also going to be studied. Do the resulting graphs show the same statistical properties than the "real networks" that they model? And, what can we learn if we compare the Semantic Web with other "semantic" Complex Systems like WordNet? It is sure that more questions are to come.

References

[1] Berners-Lee, T.: Weaving the Web. HarperBusiness (2000).

[2] Wolf, Y., Karev, G. and Koonin, E.: Scale-free networks in biology: new insights into the fundamentals of evolution? *Bioessays*, **24**, 105-109 (2002).

[3] Amaral, L.A.N., Scala, A., Barthélémy, M. and Stanley, H.E.: Classes of small-world networks. *Proc. Natl. Acad. Sci*, **97**, 11149-11152 (2000).

[4] Montoya J.M. and Solé, R.V.: Small World Patterns in Food Webs. *Theoretical Biology*, **214**, 405-412 (2002).

[5] Albert, R. and Barabási, A.L.: Statistical mechanics of complex networks. *Reviews of Modern Physics*, **74**, 47-97 (2002).

[6] Adamic, L.A.: The Small World Web, in *Research and Advanced Technology for Digital Libraries*, ed. Abiteboul, S. and Vercoustre, A. (Lecture Notes in Computer Science, Springer-Verlag, 1999).

[7] Barabási, A.L., Dezso, Z., Ravasz, E., Yook, S.H., Oltvai, Z.: Scale-free and hierarchical structures in complex networks, in *Statistical Mechanics of Complex Networks*, ed. Pastor-Satorras, R.; Rubi, J. and Diaz-Guilera, A. (Lecture Notes in Physics, Springer-Verlag, 2003).

[8] Pool, I. and Kochen, M.: Contacts and influence. *Social Networks*, **1**, 1-48 (1978).

[9] Milgram, S.: The small world problem. *Psychology Today*, **2**, 60-67 (1967).

[10] Solé, R.V., Ferrer, R., Montoya, J.M. and Valverde, S.: Selection, tinkering and emergence in Complex Systems. *Complexity*, **8**(1), 20-33 (2002).

[11] Kleinberg, J. and Lawrence, S.: The Structure of the Web. *Science*, **294**, 1849–1850 (2001).

[12] Sigman, M. and Cecchi, G.A.: Global organization of the Wordnet lexicon. *Proc. Natl. Acad. Sci*, **99**(3), 1742-1747 (2002).

[13] Ferrer, R. and Solé, R. V.: The small world of human language. *Proc The Royal Society*, **268**, 2261-2265 (2001).

[14] Gil, R.; García, R. and Delgado, J.: An interoperable framework for IPR using web ontologies, in *Legal Ontologies and Artificial Intelligence Techniques*, ed. Biasiotti, M.; Francesconi, E.; Sagri, M. and Lehman, J. (IAAIL Workshop Series, Wolf Legal Publishers, 2005).