Ongoing Research Column: Real World SW Cases

The Research in Progress or Ongoing Research Column is dedicated to the presentation of interesting research works with important achievements and critical milestones towards the realization of Information systems that prove the value potential of semantic web. In this issue we present one short article.

Measuring the Semantic Web

Rosa Gil, Roberto García, Jaime Delgado
Universitat Pompeu Fabra (UPF), Departament de Tecnologia,
Pg. Circumval·lació 8, E-08003 Barcelona, Spain
{rosa.gil,roberto.garcia,jaime.delgado}@upf.edu

Introduction

The book *Weaving the Web* by Tim Berners-Lee [i] 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 (CSs) 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 CSs are metabolic networks [i] acquaintance networks [ii], food webs [ii] or neural networks [ii]. Actually, it has also been shown that the WWW is a CS [ii].

What do all this systems have in common? How to identify a CS? 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.

Modelling and Analysing Complex Systems

Graphs are used to model CSs in order to analyse them. Nodes represent the CS 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 CSs. Some of them are highlighted here. They are considered sufficient conditions for identifying a CS:

- **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 [ii]. 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 [ii,iii], 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 [ii], $d \approx d_{random}$.
- **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, CSs show a high clustering compared to random graphs, $C >> C_{random}$. A high clustering confirms small-worldness.

Is the Semantic Web a Complex System?

We are now going to study the Semantic Web as a CS. 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 CS 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 [ii].

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 CSs, 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 stages focusing 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 CSs. For instance, FOAF metadata models social networks that have been extensively studied as CSs [viii].

Therefore, in order to collect the semantic web ontologies that are analysed, an RDF crawler is launched over the DAML Ontology Library [ii]. The processed URIs are combined in a RDF graph built from 160,000 triples.

Graph analysis

In order to analyse the obtained Semantic Web graph we use Pajek [ii], 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 Pajek network has 56,592 nodes and 131,130 arcs. Once loaded in Pajek, the available tools are used to obtain the required information about the graph: average degree, average path length, clustering factor and degree distribution.

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 CSs networks: the results from some WWW studies [ii,vi], WordNet [ii] and human language words networks [ii].

Table 1. Some CS 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	56592	4.63	0.152	4.37	-1.48
WWW	~200 M		0.108	3.10	-2.24

WordNet	66025		0.060	7.40	-2.35
WordsNetwork	500000		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 DAMLOntos is shown in Fig. 3. The linear regression of this function gives an exponent γ = -1.485 with a regression error $\varepsilon_{\%}$ = 1.455.

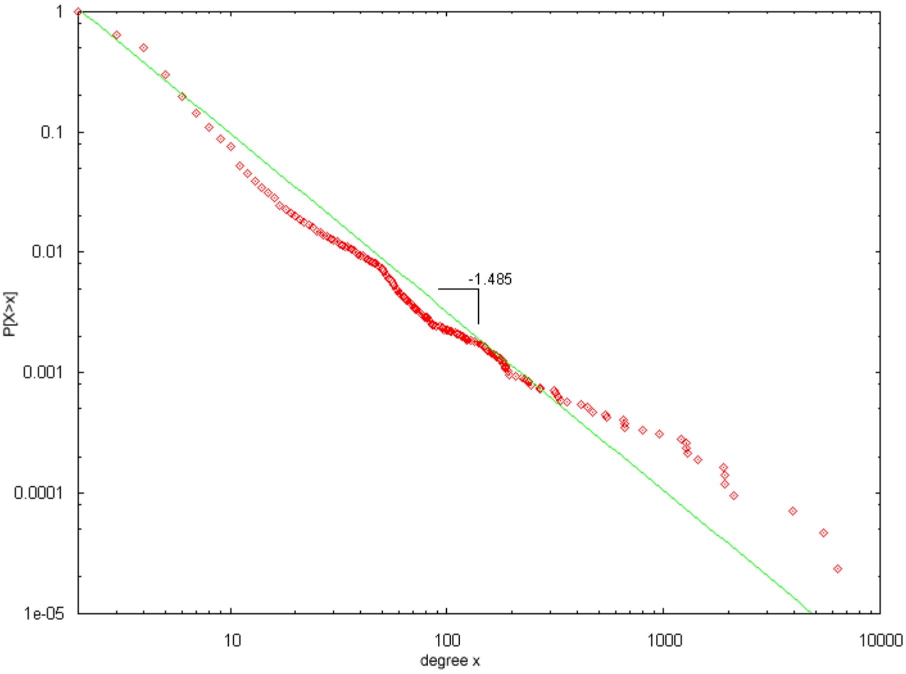

Fig. 3. Degree CDF (Cumulative Distribution Function) for the set of studied DAML library ontologies (DAMLOntos) plus linear regression and computed exponent

Therefore, the graph for the portion of the Semantic Web that has been analysed shows clear evidences that the Semantic Web behaves like a CS. 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 [i] ontology: C = 0.071 while C_{rand} = 0.0034272, <d> = 3.99 while $<d>_{rand}$ = 5.38 and γ = -1.06 with $\varepsilon_{\%}$ = 4.45.

Conclusions and future work

It has been shown that the Semantic Web behaves like a CS. When it is viewed as a graph, it reproduces all their characteristic patterns that. 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 CSs.

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" CSs like WordNet? It is sure that more questions are to come.

References

(i). Berners-Lee, T.: Weaving the Web. HarperBusiness (2000)

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

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

(iv). Montoya J.M., Solé, R.V.: Small World Patterns in Food Webs. Journal of Theoretical Biology, 405-412 (2002)

(v). Albert, R., Barabási, A.L.: Statistical mechanics of complex networks. Reviews of Modern Physics, 74 (2002) 47-97

(vi). Adamic, L.A.: The Small World Web. Proceedings of ECDL'99, LNCS 1696, Springer-Verlag (1999) 443-452

(vii). Barabási, A.L., Dezso, Z., Ravasz, E., Yook, S.H., Oltvai, Z.: Scale-free and hierarchical structures in complex networks. To appear in Sitges Proc. on Complex Networks (2002)

(viii). Pool, I., Kochen, M.: Contacts and influence. Social Networks 1 (1978) 1-48

(ix). Milgram, S.: The small world problem. Psychology Today 2 (1967) 60-67

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

(xi). Living Semantic Web project web page, http://dmag.upf.es/livingsw

(xii). DAML Ontology Library web page, http://www.daml.org/ontologies

(xiii). Pajek, http://vlado.fmf.uni-lj.si/pub/networks/pajek

(xiv). Kleinberg, J., Lawrence, S.: The Structure of the Web. Science, Vol. 294 (2001) 1849–1850

(xv). Sigman, M., Cecchi, G.A.: Global organization of the Wordnet lexicon. Proc. Natl. Acad. Sci, vol. 99, no. 3 (2002) 1742-1747

(xvi). Ferrer, R., Solé, R. V.: The small world of human language. Proc The Royal Society 268 (2001) 2261-2265

(xvii). Delgado, J., Gallego, I., Garcia, R., Gil, R.: An ontology for Intellectual Property Rights: IPROnto. Extended poster abstract, International Semantic Web Conference (2002) http://dmag.upf.es/ontologies/ipronto