The Scale-Free Nature of Semantic Web Ontology

Hongyu Zhang
School of Software
Tsinghua University
Beijing 100084, China
86-10-62773275
hongyu@tsinghua.edu.cn

ABSTRACT
Semantic web ontology languages, such as OWL, have been widely used for knowledge representation. Through empirical analysis of real-world ontologies we discover that, like many natural and social phenomenon, the semantic web ontology is also "scale-free".

Categories and Subject Descriptors
D.2.8 [Software Engineering]: Metrics—complexity measures; I.2.4 [Knowledge Representation Formalisms and Methods]: Representation languages

General Terms
Measurement.

Keywords
Scale-free, Ontology, Semantic Web, Power-law.

1. INTRODUCTION
It is widely believed that Semantic Web ontologies provide a solution to the knowledge management and integration challenges. The ontology languages such as RDF, DAML+OIL and OWL can serve as universal modeling languages for knowledge representation. A great deal of efforts is being invested in using Semantic Web ontologies to create mutually agreeable and consistent vocabularies to describe terminology and data from disparate sources [1]. For example, the NCI Thesaurus Ontology developed and actively maintained by the National Cancer Institute is an OWL ontology. It defines 60,000+ named classes, a roughly equal number of anonymous classes and 100,000+ connections (properties) from and to these classes. The OpenGALEN project also created biomedical ontologies with more than 35,000 concepts involved.

It is natural to consider a semantic web ontology as a large network, where nodes are entities (classes and individuals) and links are relationships among entities. Formally, an ontology can be viewed as a graph $G=<N, E>$, where N is a set of nodes representing entities, and E is a set of edges representing properties that link nodes. These properties include both OWL properties (such as owl:subclassOf, owl:equivalentClasses, etc.) and user-defined properties. As an example, Figure 1 shows the network view of the ProPreO ontology, which describes Proteomics data and process and is developed by the National Center for Research Resources (NCRR). In this paper, we show that a large ontology network such as the one shown in Figure 1 is "scale-free".

WWW 2008, April 21–25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

2. DISCOVERING THE SCALE-FREE NATURE OF ONTOLOGY
It is discovered that many complex networks, such as the Internet, WWW, scientific citations, protein interactions or language networks, are scale-free [2, 3]. The term "scale-free" comes from the fact that the structure and dynamics of such networks are independent of the scale of the networks. The distinguish characteristics of scale-free networks is that nodes in such networks are not randomly or evenly connected. On the contrary, some nodes are highly connected, acting as "hubs" that connects the rest of the nodes. For example, a study of *S. cerevisiae* protein-protein interaction network shows that about 93% of the proteins have five or fewer connections, while only 0.7% of the proteins have more than 15 connections [4]. More specifically, the degree distribution of nodes in scale-free networks follows the power-law, such that the probability that a node is connected to k other nodes is proportional to:

$$p(k) = C\,k^{-a}$$

where C is a constant and a is the exponent of the power-law. Taking the logarithm on both sides of the above equation, we get $ln(p(k)) = ln(C) - aln(k)$. So a power-law distribution is seen as a straight line on a log-log plot. The slope of the line is $-a$ and the intercept is $log(C)$.

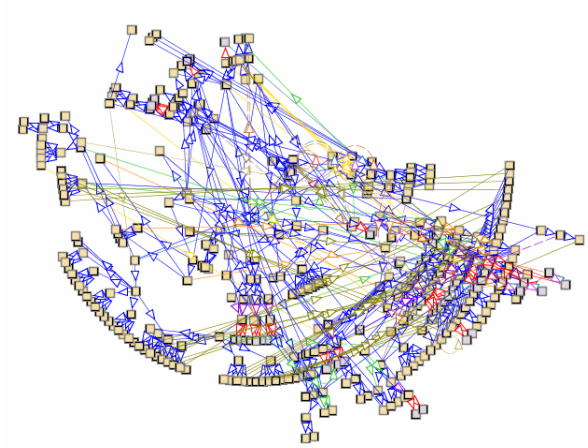

Figure 1. The ontology network of the ProPreO ontology.

To check if ontology network such as the one shown in Figure 1 is scale-free, we have collected a set of real-world biological and biomedical ontologies (as shown in Table 1). To facilitate automated data collection, we have also developed a tool, which traverses an ontology network, collects & stores relevant

information. We then analyze the degree distribution of the ontology networks.

Figure 2 shows the degree distributions of the ontology Full-Galen and NCI-Ontology. The distributions form straight lines in log-log diagrams, revealing the power-law behavior. Table 1 shows the best fit power-law parameters for all studied ontologies. The exponent *a* ranges from 2.12 to 2.47. The corresponding R^2 ranges from 0.91 to 0.99, indicating good fitness of the data (at the significant level 0.00).

Table 1. The characteristics of studied ontologies

Ontology	Description	Size (KB)	Nodes	Links	*a*	R^2
CL	An ontology for cell types	784	864	2598	2.12	0.91
Full-Galen	The full GALEN ontology of biomedical terms, anatomy and drugs translated into OWL	20100	23142	118373	2.47	0.99
Gene	The Gene Ontology project, which provides a controlled vocabulary to describe gene and gene product attributes in any organism.	39200	24316	63828	2.42	0.98
MGED	A biomaterial ontology for microarray experiments in support of MAGE	556	234	872	2.19	0.98
NCI-Ontology	The National Cancer Institute thesaurus, distributed as a component of the NCI Center for bioinformatics caCORE distribution.	32800	27653	93617	2.40	0.98
ProPreO	A comprehensive Proteomics data and process provenance ontology	229	597	1188	2.47	0.96
Tambis	A biological science ontology developed by the TAMBIS project	214	393	1732	2.36	0.99

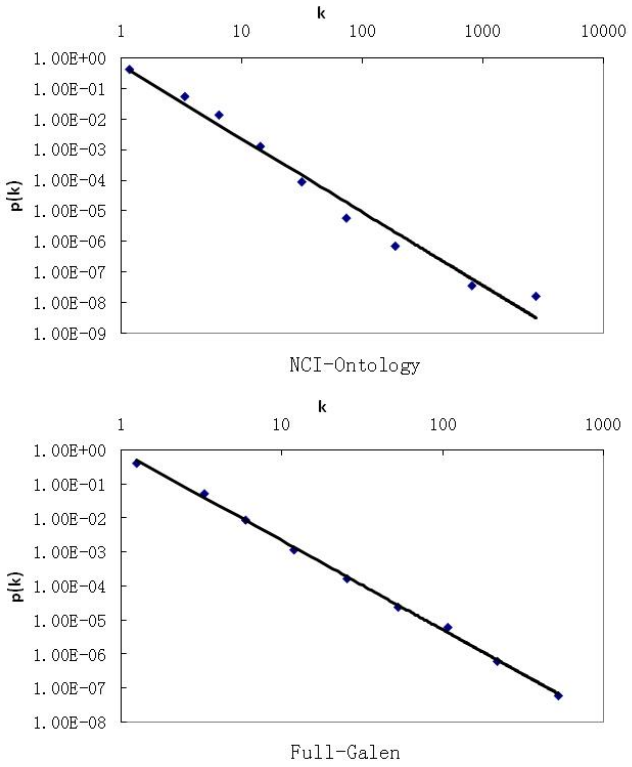

Figure 2. The power-law degree distribution of ontology network (the data is logarithmically binned).

3. CONCLUSIONS

In this short note we show that, like many natural and social phenomenon, the semantic web ontology is "scale-free". We derive the findings through empirical analysis of the degree distribution of ontology networks.

The power-law distribution of degrees implies that the knowledge represented by semantic web ontology can be very inhomogeneous: while the majority of concepts use (refer to) a few other concepts, a small number of concepts use (refer to) a large number of other concepts. The concepts that have large degrees could be treated as more "essential" knowledge points, as they attract more connections. More efforts are probably needed to understand and learn them. During maintenance, special cares need to be taken when changes to these concepts are made, as the changes may be propagated to a large proportion of the ontology.[1]

REFERENCES

[1] Ashburner, M. et al., Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature Genet*, 25(1):25–29, 2000.

[2] Barabasi, A. L. and Albert, R., Emergence of scaling in random networks. *Science* 286(5439):509.512, 1999.

[3] Barabasi, A. L. and Bonabeau, E., Scale-Free Networks, *Scientific American*, 288, 60-69, 2003.

[4] Jeong, H., Mason, S.P., Barabasi, A.L., Oltvai, Z.N., 2003. Lethality and Centrality in Protein Networks, *Nature* 411.

[1] This research is supported by the NSF China grant 60703060, and the National 863 Project 2007AA01Z122. The author thanks Dr Li Yuan Fang for the data collected.