

Semantic Interoperability and Information Fluidity

Guofei Jiang and George Cybenko
Institute for Security Technology Studies and
Thayer School of Engineering
Dartmouth College, Hanover, NH 03755
{gfj,gvc}@dartmouth.edu

James A. Hendler
Department of Computer Science
University of Maryland
College Park, MD 20742
hendler@cs.umd.edu

ABSTRACT

Ontologies are developed to describe data semantics on the Semantic Web. Given the distributed nature and scale of the Semantic Web, a large number of ontologies with different terminologies and structures will be created to describe the same concepts and domains. Without semantic mapping, information fluidity within the Web could be blocked at the boundaries of these ontologies. Therefore, ontology mapping is needed to translate datasets represented by disparate ontologies. We believe that over time communities will incrementally build an ontology mapping between select ontologies based on their own communication interests. How will these interest-driven mapping activities eventually change semantic interoperability and information fluidity across the Web? This paper proposes metrics to quantify information fluidity and builds an analytical model with “small-world” graph theory to analyze the growth of the Semantic Web. Further with this model, we analyze how information fluidity can evolve by “market-based” semantic mapping activities occurring across the Web. One conclusion, based on this model, is that the development of decentralized ontology mappings can lead to significant information fluidity within the Semantic Web.

Keywords

Distributed information systems, Semantic interoperability, Semantic web, Ontology mapping, Graph theory

1. INTRODUCTION

The immense success of the World Wide Web has dramatically changed the way that we share information. However, most information on the Web is designed for human consumption. Machines are not able to understand and process this information. For the Web to reach its full potential, it must evolve into an information space where data can be shared and processed by software agents as well as by people. The Semantic Web [18] activity centers on the infusion of meaning into the Web so as to make more of the information of the Web “machine-understandable”. The challenge of the Semantic Web is to provide languages that can allow mechanical reasoning about Web-based data and other resources [6].

An ontology is an explicit specification of the conceptualization of a domain that defines a common vocabulary to represent shared knowledge in a community of interest. Ontological commitments are agreements to use the shared vocabulary in a coherent and consistent manner [13]. Every information system in this community agrees to share the same definition of common concepts and there is thus semantic interoperability among systems that commit to the same ontology. We say that two information systems are semantically interoperable if they can understand and process semantics of their exchanged data. However, increasing the size and scope of such a community can make a centralized ontology rapidly unmanageable. Given the distributed nature and scale of the Semantic Web, it is inevitable that a large number of ontologies will be independently developed to describe the similar concepts and domains. Different terminologies and structures will be used in different ontologies to represent the same concepts [14]. Information processing across ontologies is impossible without knowing some form of semantic mapping between disparate ontologies. Therefore information sharing across the Web could be blocked at the boundaries of these ontologies.

Ontology mapping can translate information represented using one ontology to information represented in another. Therefore information systems committing to different ontologies can gain semantic interoperability between them via ontology mapping. It seems likely that communities will be motivated to build semantic mappings between various ontologies based on their own communication interests. We call this “market driven” ontology mapping – it will proceed in a self-motivated, distributed and decentralized way (motivated by the “market” needs of the users). It should be noted that the results in this paper do not assume a common ontology language, although the advent of languages such as OWL, make this market-driven model more probable in practice, since this allows greater interoperability and eases the burden of creating mappings.

The success of the Semantic Web partially relies on whether these interest-driven semantic mapping activities will eventually change semantic interoperability and information fluidity across the Web. In this paper, we propose metrics to quantify information fluidity and build an analytical model with “small-world” graph theory to analyze the growth of the Semantic Web. Further with this model, we analyze how the information fluidity will be changed by market-driven ontology mapping activities occurring across the Web.

The remainder of the paper is organized as follows: In Section 2, we discuss ontology mapping technologies and our assumptions. Section 3 formulates the semantic interoperability issue of the Web as a random graph connectivity problem. In Section 4, we define metrics to measure the information fluidity of a network. In Section 5 and 6, we analyze the information fluidity in both organized ontology mapping networks and market-driven ontology mapping networks, respectively. We discuss the limitations of our model in Section 7. Section 8 introduces the related work in this area.

2. ONTOLOGY MAPPING

In a small community, it’s not difficult for all users to agree on a common ontology. Implemented information systems can communicate with each other efficiently due to their commitment to the common ontology. In a networked environment with a very large number of information systems, such as the Semantic Web, we cannot expect that all systems will share a single centralized ontology, even within a same domain. Instead, it’s far more likely that different communities will incrementally define different terminologies and structures to represent similar concepts based on their own interests and assumptions. To understand data marked up by independently developed ontologies, semantic mappings between ontologies will be inevitable.

The differences between two ontologies representing similar concepts can include syntactic differences as well as semantic differences. Ontology mapping or translation has to consider many dimensions of mismatch between ontologies such as syntax, vocabulary, expressiveness, modeling conventions, model coverage and granularity, and representation paradigm [8]. In general, semantic mapping is much harder than syntactic translation and it needs to understand the meaning of vocabularies and their relationships. Two strategies have been used in previous work to map ontologies: One is to map a source ontology to a target ontology via one big, centralized ontology that servers as an interlingua. The other strategy is to map one ontology into another directly. (Ontolingua [3] and OntoMorph [8] are typical examples for the above two cases respectively [12].) For the results we propose in this paper, we do not need to distinguish these approaches, although we note that the latter (multiple ontologies with many mappings) is more likely and scales better than a centralized mapping system.

Mapping the relationship between two ontologies can be defined declaratively, using a set of mapping rules, or procedurally using some sort of program which inputs terms in one ontology and outputs terms in the other. Currently, mapping rules or programs have to be written manually by domain experts since there is no general technology to allow machines to understand the meaning of terminologies and their relationships. Some researchers are developing tools such as GLUE [11] to semi-automate the mapping process with machine learning technology. Once mapping rules or programs are available between two ontologies, any datasets represented with one ontology can automatically be mapped to datasets represented with another ontology. Two ontologies may only have partial mapping between their terminologies and structures because of many dimensions of mismatches listed above. The mapping could be implemented to work only in one direction but not in the reverse direction. However, we believe that ontology mapping is likely to be a bilateral collaboration of developers at both sides because an accurate mapping needs developers to have thorough understanding on both ontologies. For the communication needs of both sides, developers tend to build mappings bi-directionally with a little extra work. We will discuss more about this in Section 6.2. In this paper, we try to get a macroscopic view over the information fluidity across the Semantic Web so that we make the simplifying assumption that all mappings are invertible (that is, designed to be bi-directional).

3. ONTOLOGY AND SYSTEM NETWORK

It is useful to view ontologies and their mapping relationships as a complex ontology network. Traditionally topologically complex networks have been described using random graph theory. Here, each ontology is represented with a vertex in the ontology graph. If there exists an ontology mapping between two ontologies, the associated two vertices are linked with an edge. Otherwise, there is no edge between these two vertices. Since we assume that the ontology mapping is bi-directional, the ontology graph is an undirected graph. If two vertices are connected with a path of consecutive edges, we say that these two nodes are connected in the graph. Further, if every pair of nodes in the graph is connected, we say that the graph is fully connected. If a graph is not fully connected, the fully connected sub-graphs are named as components. If there exists a single component that links the majority of nodes in the graph, this component is named as the giant component.

We can also view all the information systems using ontologies on the Web, and their semantic interoperability relationships as a system network. Every information system employs one ontology from the ontology network to markup its data (Note that in practice a system might use multiple ontologies to represent different views of its data. For the analysis in this paper, we treat such a system as if it is multiple systems, each linked to a single ontology, with no loss of generality in our analysis). If we cluster those systems committing to the same ontology together, the whole information system network on the Web can be partitioned into many clusters. Each cluster can be mapped into a vertex in the ontology graph. On the other side, from the ontology network's view, every system is a markup instance of a specific ontology and each cluster is a set of markup instances of a specific vertex in the ontology graph. The relationship between the ontology network and the system network is shown in Figure 1. In the ontology graph, the edge between two vertices represents that there exists an ontology mapping between two associated ontologies. In the system graph, the edge between two nodes represents that the associated two systems are semantically interoperable. The edge between two clusters represents that the nodes in these two clusters are semantically interoperable via ontology mapping, i.e. every pair of nodes from these two clusters have semantic interoperability between them. Since the systems in the same cluster commit to the same ontology, they're semantically interoperable directly and therefore the nodes in the same cluster are fully meshed.

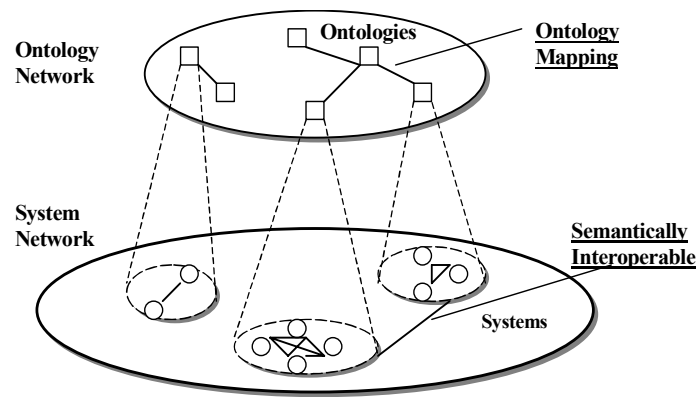


Figure 1: The relationship between ontology and system network

For the Semantic Web to reach the Internet scale, ontology-mapping resources have to act analogously to the routers in the Internet. While large numbers of nodes share backbone routers to achieve global network connectivity, information systems have to cooperatively share ontology mapping resources to achieve global semantic interoperability. Two ontologies in the ontology network may only have partial mapping between them. Various applications of an information system may demand different mapping accuracies to process the data represented in other ontologies. While some critical applications may need 100% mapping accuracy to process data, others may only need partial mapping to extract the required portion of information. In fact, one ontology mapping could consist of several partial mappings, i.e. One ontology can be split to several parts, which are independently mapped to the target ontology via different mapping paths in the ontology graph. Therefore partial mapping could also satisfy the interoperability needs of some information systems. Like in the Internet, one message could be split to several packets and these packets could be routed via different paths to the destination, which assembles these packets back into one message. In this paper, we focus our analysis on the information fluidity or reachability that results from

mapping connectivity. Whether an ontology mapping is acceptable should be determined by individual applications. Like the OSI network model, we may need several layers to address different issues arising in ontology mapping. It seems that mapping quality control mechanism should be addressed at the application layer. At the mapping connectivity layer, we can assume that all edges in the ontology graph have roughly the same mapping quality.

4. INFORMATION FLUIDITY

If two nodes are semantically interoperable in the system network, we say that information can flow from one node to the other and vice versa. For the sake of this paper, we will abstract away the details of how the interoperability works – the analysis is the same whether we assume wrappers, web services, agents or any other mechanisms that allow for loosely-coupled distributed computing across network. Our model will analyze how information fluidity of the Web is affected by the semantic interoperability of distributed information systems.

In a network of information systems, we use the ratio between the largest number of nodes that are fully connected with regard to semantic interoperability and the total number of nodes in the network as a metric to quantify the information fluidity of that network. For example, every cluster that commits to the same ontology has 100% information fluidity within that cluster since the cluster is fully meshed with regard to semantic interoperability. Though there possibly exist many fully connected components in a system network after ontology mapping, we only use the size of the largest component to represent the information fluidity of that network. If more nodes in the network are semantically interoperable, the network is claimed to have better information fluidity. Straightforwardly, ontology mapping can dramatically improve the information fluidity in a heterogeneous Web since it connects different clusters in the system network. For convenience in the following sections, here we prove several simple propositions:

Proposition 1: If two vertices in the ontology graph are connected, the associated two clusters in the system network are semantically interoperable.

Proof: If two vertices in the ontology graph are connected, there exists at least one path with consecutive edges between these two nodes in the ontology graph. The two ontologies represented by these two nodes can be mapped into each other via a sequence of ontology mappings, which are represented by the sequence of edges along the path. Therefore every pair of nodes in the associated two clusters is semantically interoperable via this sequence of ontology mappings. ■

Here we assume that any length of mapping sequence between two nodes is acceptable. In reality, a long sequence of ontology mappings may lead to large computing overhead, low mapping accuracy and/or high latency in data processing. However, as we will see in section 6, in small-world phenomenon, nodes are usually connected with a short chain of edges. The small-world phenomenon is colloquially called “six degrees of separation”, i.e., typically only six edges between two connected nodes in a massive network.

Proposition 2: If the ontology graph is fully connected, every pair of nodes in the system network is semantically interoperable and the whole information system network has 100% information fluidity.

Proof: If the ontology graph is fully connected, by definition every pair of nodes in the ontology graph is connected. According to Proposition 1, every pair of associated clusters in the system network is semantically interoperable via a sequence of ontology mappings. That means every pair of nodes from any two clusters is semantically interoperable indirectly via a sequence of ontology mappings. Since all nodes within the same cluster commit to the same ontology, they are semantically interoperable with each other directly. Therefore, every pair of nodes in the system network is semantically interoperable directly or indirectly (via ontology mapping). The whole information system network is fully connected with regard to semantic interoperability. According to our metric definition of information fluidity, 100% of nodes in the information system network are semantically interoperable and the whole network has 100% information fluidity. ■

5. ORGANIZED ONTOLOGY NETWORKS

According to Proposition 2, if the ontology graph is fully connected, then the system network has 100% information fluidity. While this could be an ideal situation for the system network with regard to information fluidity, some interesting questions remain: Assuming that global ontology mapping activities can be well organized, what is the

minimum effort to get the ontology graph fully connected without loops? Conversely what is the worst case to get the ontology graph fully connected?

Proposition 3: In an ontology network with M ontologies, at least $M-1$ ontology mappings and at most $\frac{(M-1) \cdot (M-2)}{2} + 1$ ontology mappings are needed to get 100% information fluidity in the associated information system network.

Proof: For a graph with M vertices, straightforwardly at least $M-1$ edges are needed to get the graph fully connected (best case). Conversely, in the worst case, it could take $\frac{(M-1) \cdot (M-2)}{2} + 1$ edges to get a graph with

M vertices fully connected, i.e., $M-1$ vertices are fully meshed before a new edge gets the last isolated vertex connected. According to Proposition 2, with a fully connected ontology graph, the associated information system network has 100% information fluidity. ■

If M is a very large number, it could take significant amount of work to build $M-1$ ontology mappings and get the whole information system network connected. It seems probable, however, that in practice there will be dominant ontologies in a network that are much more popular than others. Most information systems may use these ontologies to describe their data and these ontologies have large clusters in the system network. Therefore, even if we only build ontology mappings between these dominant ontologies, the information system network could still achieve great information fluidity. For example, if 80% of system nodes use the same m ($m \ll M$) dominant ontologies to describe their data, according to Proposition 3, we only need to build $m-1$ mappings to get these m dominant nodes fully connected and the whole information system network could still obtain 80% information fluidity.

6. MARKET-DRIVEN ONTOLOGY NETWORKS

In a small domain, ontology mapping activities could be well organized as discussed in section 5, with a careful design of minimal mappings to obtain maximal fluidity. In reality, however, it seems more likely that ontology network may grow over time, with clusters forming around ontologies built for a specific purpose, and later mappings being developed to provide greater interoperability among those being widely used. Thus, new ontology nodes are added to the network incrementally and new ontology mappings (edges) are built between existing nodes incrementally. Given the distributed nature and scale of the Semantic Web, eventually it may grow to a massive network with no truly dominant nodes. Due to its scale and complexity, it's not realistic to organize this network rigorously. Instead, the ontology network is more likely to grow based on a "market-driven" approach (in self-motivated, distributed and decentralized way), with much reuse of existing ontologies and mappings growing between those popular ones [14]. For example, if an ontology is very popular among information systems, other ontologies are more likely to build mappings with this ontology in order to achieve better information fluidity. As a consequence, this ontology may even become more popular. Recently it has been demonstrated that many large networks share certain universal characteristics that can be described by so-called the "power law" distribution [1][3]. Barabasi et al. [4] show that a power-law degree distribution and small-world phenomenon emerges naturally from a stochastic growth process in which new vertices link to existing ones with a probability proportional to the degree of the target vertex. Chung and Lu [9] analyzed random graphs with general expected degree distributions and special emphasis is given to sparse graphs with average degree a small constant. In this section, we introduce their complex graph theory first and then we apply their results to a market-driven network model built for the ontology network.

6.1 Random Graph Theory

Assume that a random graph has n nodes and a given expected degree sequences $w = (w_1, w_2, \dots, w_n)$. The vertex v_i is assigned with a vertex weight w_i that is the expected degree of this node. The edges are chosen independently and randomly according to the vertex weights as follows. The probability p_{ij} that there is an edge between v_i and v_j is proportional to the product $w_i w_j$ where i and j are not required to be distinct. There are possible loops at v_i with probability proportional to w_i^2 , i.e.,

$$p_{ij} = \frac{w_i w_j}{\sum_k w_k} \text{ and assume } \max_i w_i^2 \leq \sum_k w_k. \quad (1)$$

This assumption ensures that $p_{ij} \leq 1$ for all i and j . According to Equation 1, for a node i , its expected degree is $\sum_{j=1}^n p_{ij} = w_i$. Here we denote a random graph with a given expected degree sequence w by $G(w)$. The expected average degree of a random graph $G(w)$ is defined to be $d = \frac{1}{n} \sum_{i=1}^n w_i$. For a subset S of vertices, the volume of S , denoted by $Vol(S)$, is the sum of expected degrees in S , i.e. $Vol(S) = \sum_{v_i \in S} w_i$. In particular, the volume of $Vol(G)$ of $G(w)$ is just $\sum_{i=1}^n w_i$. With regard to a random graph $G(w)$ like this, Chung and Lu [9] proved the following theorem.

Theorem 1 (Chung and Lu 2002): For a random graph G with a given expected degree sequence having average degree $d > 1 + \delta > 1$, almost surely G has a unique giant component.

(i) If $d \geq e$, the volume of the unique giant component is almost surely at least

$$(1 - \frac{2}{\sqrt{de}} + o(1))Vol(G).$$

(ii) If $1 + \delta \leq d \leq e$, the volume of the unique giant component is almost surely at least

$$(1 - \frac{1 + \log d}{d} + o(1))Vol(G).$$

6.2 Model and Parameters

In our model, we assume that the system network has N nodes (information systems) and the ontology network has M ($M < N$) nodes (ontologies). As we discussed in section 3, every node in the system network picks one ontology from the ontology network to markup its data. Therefore the N nodes in the system network can be partitioned to M clusters. All information systems in the same cluster commit to the same ontology in the ontology network. Each cluster includes n_i ($i = 1, 2, \dots, M$) systems and $\sum_{i=1}^M n_i = N$. Define $n_{max} = \max_i n_i$, i.e. n_{max} is the size of the largest cluster of information systems.

If one ontology has more popularity in the information system network, it will gain more visibility in the ontology network. Other ontologies are more likely to build mappings to that ontology in order to get better information fluidity for their information systems. This follows “the rich get richer” phenomenon in society (and is why we refer to this as “market-driven” approach). Therefore, the degree of an ontology node should reflect the size of its cluster in the information system network. Here we assume that the expected degree of a node i in the ontology graph, w_i , is proportional to its popularity in the information system network, i.e. the degree of an ontology node is proportional to the size of its associated cluster in the system network. We define

$$w_i = K \cdot \frac{n_i}{N} \quad i = 1, 2, \dots, M, \quad (2)$$

where K is the sum of all nodes’ degrees in the ontology graph. In fact, with Equation (2), we have

$$\sum_{i=1}^M w_i = \sum_{i=1}^M K \cdot \frac{n_i}{N} = \frac{K}{N} \cdot \sum_{i=1}^M n_i = K. \quad (3)$$

Meanwhile, as shown in Equation (1), the probability p_{ij} that there is an ontology mapping between nodes v_i and v_j is proportional to the product $w_i w_j$, i.e.

$$p_{ij} = \frac{w_i w_j}{\sum_k w_k} = K \cdot \frac{n_i n_j}{N^2} \quad \text{with } K \leq \frac{N^2}{n_{max}^2} \quad (4)$$

What does Equation (4) mean in our model? As we discussed in Section 2, building semantic mapping between two ontologies is likely a bilateral activity and developers need collaboration from both sides to map terminologies and structures in both ontologies. If two ontology nodes are both popular (with big clusters in the information system network), they are both “attractive” to each other and more likely to cooperate in building a mapping between them for greater information reachability. Otherwise, these big clusters are unable to consume large amount of information from each other in a global network of information systems.

In our model, the expected degree $d = \frac{1}{M} \sum_{i=1}^M w_i = \frac{K}{M}$ and according to Equation (2), the volume of subset S is

$$Vol(S) = \sum_{v_i \in S} w_i = \frac{K}{N} \cdot \sum_{v_i \in S} n_i. \quad (5)$$

The volume of the whole ontology network G is

$$Vol(G) = \sum_{i=1}^M w_i = \frac{K}{N} \cdot \sum_{i=1}^M n_i = K. \quad (6)$$

6.3 Lower Bound of Information Fluidity

With the above model and parameters, we can derive from Theorem 1 the following lower bound of information fluidity in a network of information systems.

Theorem 2: In an ontology network $G(M, L)$ with M ontology nodes and L ($L \leq \frac{N^2}{2n_{max}^2}$) ontology mapping links, if the edge number of an ontology vertex is proportional to the number of information systems using this ontology, then

- (i) If $L \geq \frac{M \cdot e}{2}$, the information fluidity of the whole information system network is almost surely at least

$$1 - \sqrt{\frac{2M}{L \cdot e}} + o(1).$$

- (ii) If $\frac{M(1+\delta)}{2} \leq L \leq \frac{M \cdot e}{2}$, the information fluidity of the whole information system network is almost surely at least

$$1 - \frac{M + M \cdot \log \frac{2L}{M}}{2L} + o(1).$$

Proof: In a random graph, the sum of all nodes’ degrees is always twice of the number of edges. In an ontology network $G(M, L)$ with M nodes and L mapping links, the sum of degrees is: $K = 2L$. According to the constraints in Inequalities (1) and (4), $L = \frac{K}{2} \leq \frac{N^2}{2n_{max}^2}$, i.e. only those networks with spare links can apply this

result. The average degree of these M nodes is: $d = \frac{K}{M} = \frac{2L}{M}$. According to Theorem 1, if $d > 1 + \delta > 1$, almost surely $G(M, L)$ has a unique giant component. Denote this giant component as S_{giant} . Moreover, if $d = \frac{2L}{M} \geq e$, i.e. $L \geq \frac{M \cdot e}{2}$, the volume of this unique giant component is almost surely

$$Vol(S_{giant}) \geq (1 - \frac{2}{\sqrt{de}} + o(1))Vol(G). \quad (7)$$

Therefore with $d = \frac{2L}{M}$, we have

$$\begin{aligned} \frac{Vol(S_{giant})}{Vol(G)} &\geq \frac{\left(1 - \frac{2}{\sqrt{de}} + o(1)\right)Vol(G)}{Vol(G)} \\ &= 1 - \frac{2}{\sqrt{\frac{2L \cdot e}{M}}} + o(1) \\ &= 1 - \sqrt{\frac{2M}{L \cdot e}} + o(1). \end{aligned} \quad (8)$$

Meanwhile, by the definition of $Vol(S)$ and $Vol(G)$ in Equation (5) and (6) respectively, we have

$$\frac{Vol(S_{giant})}{Vol(G)} = \frac{\frac{K}{N} \cdot \sum_{v_i \in S_{giant}} n_i}{K} = \frac{\sum_{v_i \in S_{giant}} n_i}{N}. \quad (9)$$

By Inequality (8) and Equation (9), we have

$$\frac{\sum_{v_i \in S_{giant}} n_i}{N} \geq 1 - \sqrt{\frac{2M}{L \cdot e}} + o(1). \quad (10)$$

The giant component S_{giant} is the largest sub-graph that is fully connected in the ontology network. For any vertex $v_i \in S_{giant}$ in the ontology network, n_i nodes in the system network commit to that ontology. Therefore, a total of $\sum_{v_i \in S_{giant}} n_i$ nodes in the system network commit to the ontologies that are included in the giant component S_{giant} . Since the giant component S_{giant} is fully connected, according to Proposition 2, any pair of these $\sum_{v_i \in S_{giant}} n_i$ nodes is semantically interoperable, directly or indirectly (via ontology mapping). Therefore in the information system network, the largest percentage of nodes that are semantically interoperable between each other is at least $\frac{\sum_{v_i \in S_{giant}} n_i}{N}$. By the metric definition of information fluidity and Inequality (10), we can conclude that the information fluidity of the information system network is almost surely at least $1 - \sqrt{\frac{2M}{L \cdot e}} + o(1)$.

In the similar way, if $\frac{M(1+\delta)}{2} \leq L \leq \frac{M \cdot e}{2}$, we can conclude the result (ii) from the second part of Theorem 1. ■

6.4 Results Analysis

In this section, we use an example to illustrate Theorem 2. Assume that the ontology network has $M = 5000$ ontologies and the largest cluster in the system network includes 0.5% of information systems, i.e. $\frac{n_{max}}{N} = 0.005$.

Given the scale of the Web, N could be a very large number. Without ontology mapping, the information fluidity of the system network is only 0.5% according to the metric definition of information fluidity, i.e., these $0.5\% \cdot N$ nodes are fully connected in the system network with regard to semantic interoperability. According to the constraints in

Inequalities (1) and (4), we have $L \leq \frac{N^2}{2n_{max}^2} = 20000$. As we mentioned earlier, our theorem only applies to those

networks with sparse edges. In this specific example, it only applies to the network with edges less than 20000. Assume that we have $L = 10000$ ontology mappings that are randomly distributed according to Equation (4). By the result (i) of Theorem 2, the information fluidity of the information system network is at least 40%. That is, after 10000 self-motivated, distributed and decentralized ontology mappings, at least $40\% \cdot N$ information systems are semantic interoperable with each other.

With a network of 5000 ontologies, Figure 2 illustrates how the lower bound of information fluidity is improved by adding more ontology mappings. Given a number of ontology mappings, we can consult the curve to estimate the information fluidity of an information system network. Conversely, given a requirement of information fluidity, we can estimate the number of ontology mappings needed in an ontology network. Therefore we believe that our results can be useful in evaluating mapping efforts needed for large-scale heterogeneous information systems. Given different sizes of ontology networks, Figure 3 shows the growth of information fluidity as the number of ontology mappings increases. Given different number of ontology mappings, Figure 4 illustrates how the information fluidity decreases as the size of ontology networks grows.

Given a fixed number of ontology mappings, if the size of ontology network is bigger, the information fluidity of its system network is lower. Straightforwardly, this is because the same number of ontology mappings is more widely distributed in a bigger size network and averagely each ontology node has smaller degrees, which leads to lower connectivity in the ontology network and further lower information fluidity in the system network. In Figure 3, we note that the lower bound of information fluidity grows faster for smaller M . For example, the curve with $M = 1000$ has much steeper growth than any of others in the early stage. With the same rate of mapping growth, averagely the degree of each node in a smaller network should grow faster than in a bigger network. Therefore the probability to link two nodes increases more significantly in a smaller network according to Equation (4), which leads to the faster growth of ontology network connectivity and further the faster growth of information fluidity in the system network.

In Figure 3, we also note that after a fast growth, the information fluidity increases slowly with the growth of ontology mappings. For example, for an ontology network with 1000 nodes, i.e., $M = 1000$, with 3000 mappings, the information fluidity of its system network could reach 55%. However, with extra 7000 mappings, its information fluidity only increases 15%. Ontologies with big clusters in the system network have priority to get connected with each other, which leads to a fast growth of information fluidity. After these ontologies are fully connected, many mappings are assigned to build “redundant paths” between these ontologies, which doesn’t increase information fluidity. Those ontologies with small clusters in the system network are isolated based on the mapping distribution in Equation (4). However, as we see in Figure 2 and 3, though the number of mappings is much bigger than $M - 1$ in order to get high information fluidity, it is still much smaller than the number in the worst case: $\frac{(M-1) \cdot (M-2)}{2} + 1$. For $M = 1000$, this number is as high as half million.

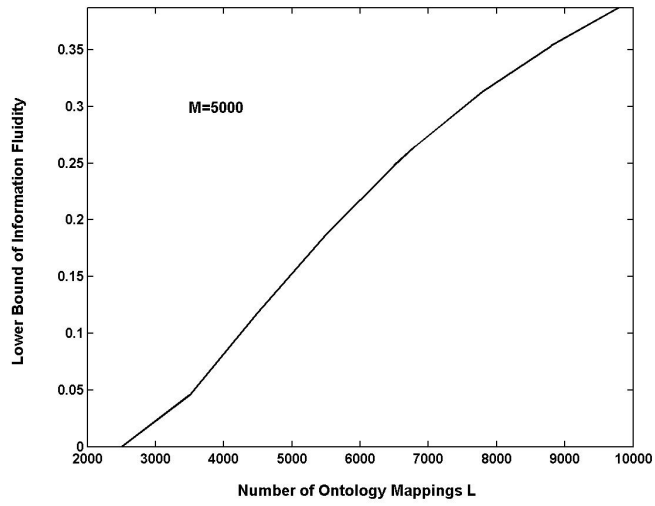


Figure 2: Information fluidity vs. number of ontology mappings ($M=5000$)

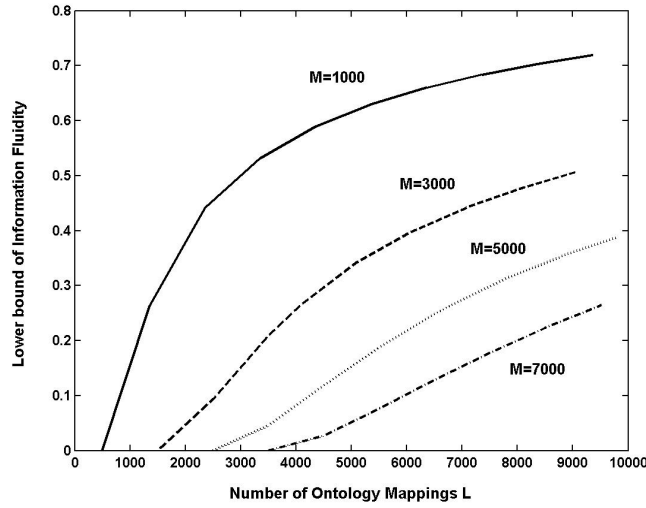


Figure 3: Information fluidity growth with different sizes of ontology networks

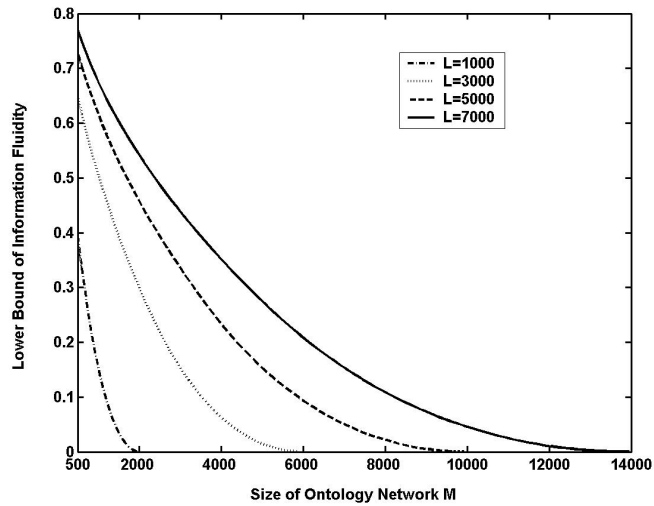


Figure 4: Information fluidity vs. size of ontology networks

7. DISCUSSIONS

For a market-driven ontology network, Theorem 2 gives the lower bound of information fluidity in its associated information system network, i.e. the worst-case scenario. That means the real information fluidity could be higher. Meanwhile, there may exist many fully connected components in the system network. All information systems within the same component have semantic interoperability between each other and each component has 100% information fluidity within its boundary. However, these components are not connected. Without ontology mapping, information fluidity is blocked at the boundary of system clusters that commit to the same ontology. With ontology mapping, it is blocked at the boundary of components. According to our metric definition for information fluidity, we use the size of the largest component to quantify the information fluidity in a system network. Therefore, the metric doesn't mean that only that percentage of system nodes have semantic interoperability. Instead, many system nodes may have semantic interoperability within their smaller component but not beyond that component.

We have to point out that our mathematical model has the following limitations, which we will address in our future work.

1. According to Equation (1), there are possible loops at vertices in the graph model illustrated in Section 6.1. In an ontology network, these loops do “waste” some number of mappings from L since they are not “valid” ontology mappings between different ontologies. Therefore the lower bound of information fluidity should be higher if these mappings are used to link different nodes in the ontology graph.
2. Our theorem only applies to those graphs with sparse links, i.e. the total number of mappings L has to be less than $\frac{N^2}{2n_{max}^2}$. However, we argue that the ontology graph of the Semantic Web will possibly be a sparse graph for a long time because the Web has such a large scale and it takes time for communities to develop mappings incrementally.

8. RELATED WORK

Inspired by empirical studies of massive networked systems such as the Internet and World Wide Web, researchers have developed various models to characterize complex networks, including the small-world model [16]. The power-law distribution has been applied to model the connectivity of nodes in many complex networks. Recently, Newman [16] and Albert et al. [3] have done comprehensive reviews of the research on complex networks. The structure of complex networks has become a popular research topic.

Much recent work has applied graph-theoretic methods to analyze the hyperlink structure of the World Wide Web. Barabasi and Albert [5] and Broder et al. [7] examine millions of Web pages in different domains and report that both the in and out degree of nodes on the Web graph follow power laws. The In-degree refers to the number of distinct links to a node. The Out-degree refers to the number of distinct links from a node. Further, Kumar et al. [15], Albert and Barabasic [3] and Aiello et al. [2] propose some stochastic models and dynamical processes to explain the random growth of the Web graph. Network edges are added to the network incrementally and they have preferential attachment to those nodes with high degrees. Recently, Dill et al. [10] show that the Web emerges as the outcome of a number of essentially independent stochastic processes that evolve at various scales. This scale invariance leads to a “fractal” structure of the Web. There exist strongly connected and weakly connected components on the Web.

The work inspired us to consider that ontology mapping activities could also follow the small-world phenomenon during the evolution of the Semantic Web. However, we believe that the problem addressed in this paper is very different. While the above work is to explain the link structure of the Web graph, we're interested in how the information fluidity could be changed by the small worlds created by market-driven semantic mapping activities across the Web. We build a two-layer model to reflect the relationship between the ontology graph and its information system graph. For a market-driven network, we introduce mapping factors to the model and produced analytical results to measure the lower bound of information fluidity. While we acknowledge that this is only a first simplified result, we believe that it is indicative that the small-world graphs may have a role in showing the potential success of the networks of ontologies discussed in [2,14] and in analyzing the advantage of an approach with many ontologies mapped to each other.

9. SUMMARY

The Semantic Web employs ontologies to represent data semantics on the Web. The ability for a machine to understand data semantics depends on the ability to share ontologies in a coherent and consistent manner. Given the distributed nature and the scale of the Web, a centralized ontology is unmanageable, even if it's possible. Instead, a large number of ontologies with different terminologies and structures will be created to describe similar concepts and domains. To understand data represented by independently developed ontologies, semantic mapping between ontologies seems to be inevitable. Otherwise information fluidity across the Web could be blocked at the boundaries of these ontologies. With ontology mappings, information fluidity is blocked at the boundary of components.

In this paper, we described a two-layer graph model that characterizes the relationship between an ontology network and a systems network. We formulized the information fluidity problem as a graph connectivity problem. For market-driven approach to ontology mapping, a "small-world" phenomenon was introduced to model interest-driven ontology mapping activities that are likely to occur when ontologies are published on the Web. Further, with this stochastic model, we presented an analytical result to compute the lower bound of information fluidity given the production of a number of ontology mappings. Based on this result, we analyzed how the information fluidity is improved with the growth of ontology mappings. Despite some limits of our approach, we have built a reasonable model that enables us to have a macroscopic view over the growth of information fluidity on the Semantic Web.

10. ACKNOWLEDGEMENTS

This research was partially supported by: Defense Advanced Research Projects Agency projects F30602-00-2-0585 and F30602-98-2-0107; the Office of Justice Programs, National Institute of Justice, Department of Justice award number 2000-DT-CX-K001 (S-1) as well as grants from NSF, NIST, ARL and Fujitsu Laboratories of America.

11. REFERENCES

- [1] Aiello, W., Chung, F., and Lu, L.. Random evolution in massive graphs. Handbook on Massive Data Sets, (Eds. James Abello et al.), Kluwer Academic Publishers, 97-122, 2002.
- [2] Aiello, W., Chung, F., and Lu, L.. A random graph model for massive graphs. In Proceedings of the 32nd STOC, 2000.
- [3] Albert R., and Barabasi, A.. Statistical mechanics of complex networks. Reviews of Modern Physics, vol. 74, January, 2002.
- [4] Barabasi, A., Albert, R., and Jeong, H.. Mean-field theory for scale-free networks. Physica A, vol. 272, 173-187, 1999.
- [5] Barabasi, A., and Albert, R.. Emergence of scaling in random networks. Science 286, 509, 1999.
- [6] Berners-Lee, T., Hendler, J., and Lassila, O.. The Semantic Web. Scientific American, May 2001.
- [7] Broder, A., Kumar, R., Maghoul, F., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J.. Graph structure in the web. In Proceedings of the 9th WWW/Computer Networks 33, 2000.
- [8] Chalupsky, H.. OntoMorph: A Translation System for Symbolic Knowledge. Principles of Knowledge Representation and Reasoning: Proceedings of the Seventh International Conference, Breckenridge, Colorado, USA, Apr. 12-15, 2000.
- [9] Chung, F., and Lu, L.. Connected components in random graphs with given expected degree sequences. Annals of Combinatorics, vol. 6, no. 2, 2002
- [10] Dill, S., Kumar, R., Mccurley, K., Rajagopalan, S., Sivakumar, D., and Tomkins, A.. Self-similarity in the web. ACM Transactions on Internet Technology, vol.2, no.3, 2002.

- [11] Doan, A., Madhavan, J., Domingos, P., and Halevy, A.. Learning to Map between Ontologies on the Semantic Web. WWW 2002, Honolulu, Hawaii, USA, May 7-11, 2002.
- [12] Dou, D., McDermott, D., and Qi, P.. Ontology Translation by Ontology Merging and Automated Reasoning. Proc. EKAW Workshop on Ontologies for Multi-Agent Systems.
- [13] Gruber, T.R.. A translation approach to portable ontologies. Knowledge Acquisition, vol. 5, no. 2, 199-220, 1993.
- [14] Hendler, J.. Agents on the Semantic Web, IEEE Intelligent Systems Journal, March/April, 2001.
- [15] Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A., and Upfal, E.. Stochastic models for the web graph. In Proceedings of the 41th FOCS Conference, 2000.
- [16] Newman, M.E.. The structure and functions of complex networks. SIAM Review, vol. 45, no. 2, 2003.
- [17] Watts, D.J., and Strogatz, S.H.. Collection dynamics of “small-world” networks, Nature, 393, 1998.
- [18] World Wide Consortium (W3C) Semantic Web Activity
<http://www.w3.org/2001/sw/>