

International Journal of Cooperative Information Systems
© World Scientific Publishing Company

SEMANTIC INTEROPERABILITY AND INFORMATION FLUIDITY

GUOFEI JIANG

*NEC Laboratories America
4 Independence Way, Princeton, NJ 08540, USA*

GEORGE CYBENKO

*Thayer School of Engineering
Dartmouth College, Hanover, NH 03755, USA*

JAMES A. HENDLER

*Department of Computer Science
University of Maryland, College Park, MD 20742, USA*

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-driven” semantic mapping activities occurring across the Web. Our results can be useful in evaluating mapping efforts needed for large-scale heterogeneous information systems. 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 systems; semantic interoperability; information fluidity; 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 Web¹ activity centers on the infusion of meaning into the Web so

2 *G. Jiang, G. Cybenko & J. Hendler*

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.²

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.³ 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 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.⁴ 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 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. We believe that our results can be useful in evaluating mapping efforts needed for large-scale heterogeneous information systems.

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.⁵ Some work on multi-database systems also classified semantic and schematic conflicts between heterogeneous database objects.^{6,7} 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. (Ontolingua³ and OntoMorph⁵ are typical examples for the above two cases respectively.⁸) 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 GLUE⁹ 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.

For two information systems using two ontologies to be semantic interoperable, it seems that bi-directional ontology mappings are needed between these two ontologies. Assume that a system s_1 using ontology o_1 needs to communicate with another system s_2 using ontology o_2 . A query from the system s_1 (composed with ontology o_1) has to be mapped into a query using terms of the target ontology o_2 so that the system s_2 can “understand” and process the query. Meantime, the data responded from the system s_2 (composed with ontology o_2) has to be mapped back into the ontology o_1 so that the system s_1 can “understand” and process the returned data. This is especially necessary if it is autonomous software agents that process query and data without human’s intervention. A software agent may need to correlate and process queries from multiple sources but only be able to process data marked up with its own ontology. In such interactions, the information flow is bi-directional and bi-directional ontology mappings are very much needed. Note that we don’t assume that all terms and their relationships of two ontologies can be mapped into each other without information loss. Two ontologies may only have partial mapping between their terminologies and structures because of many dimensions of mismatches listed above. Some terms and relationships in one ontology may not be mapped into another and in our framework such information loss is allowed during mapping. Here we only assume that in general two ontologies (but not every individual term and/or relationship) should have bi-directional mapping relationship in order to enable information systems using these ontologies to be semantically interoperable, as illustrated above. We say that there is a bi-directional mapping between two ontologies even if each ontology only has a subset of terms and/or relationships mapped into another in one direction and these two sets are not overlapped. See more discussions on mapping accuracy and information loss in Section 3. Some earlier work has proposed several approaches to resolve the conflicting terms in ontology mapping. For example, in the OBSERVER system,¹⁰ if a term in one ontology can not be translated into the target ontology, this conflicting term is replaced by the intersection of its immediate parents or by the union of its immediate children recursively until a translation of the conflicting term is obtained using only the terms of the target ontology.

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, which consists of many different ontologies. If there exist ontology mappings 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 in general two ontologies have bi-directional mapping relationship to support system interactions, 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 con-

nected 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 an information 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 information systems committing to the same ontology together, the whole information system network can be partitioned into many clusters. Each cluster of information systems is associated with one vertex in the ontology network. On the other side, from the ontology network's view, every information 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 network. The relationship between the ontology network and the information system network is shown in Figure 1. In the ontology network layer, the edge between two vertices represents that there exist ontology mappings between two associated ontologies. In the system network layer, the edge between two nodes represents that the associated two information systems are semantically interoperable. The edge between two clusters represents that the information systems in these two clusters are semantically interoperable, i.e., every pair of information systems from these two clusters is semantically interoperable. Since all information systems in the same cluster commit to the same ontology, straightforwardly they're semantically interoperable and

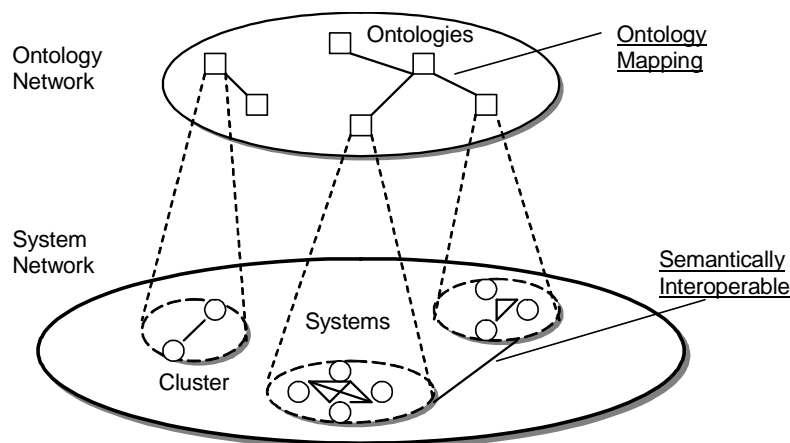


Fig. 1. The relationship between ontology and system network

therefore all nodes in the same cluster are fully meshed. Note that in the following sections, a “node” in the ontology network always refers to an individual ontology while a “node” in the system network always refers to an individual information system.

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 mappings between them, which lead to information loss during the mapping process. However various information systems may demand different mapping accuracies to process the data represented in other ontologies. While some information systems may need 100% mapping accuracy to process data, others may only need partial mappings 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 network. Therefore partial mappings could also satisfy the interoperability needs of some information systems. In this paper, we focus our analysis on the information fluidity or reachability that results from mapping activities, i.e., how information can flow in the information system network with regard to semantic interoperability. Though a partial mapping between ontologies could cause information loss, it could still enable certain amount of information (more or less) to flow from one information system to another. As discussed above, whether the quality of an ontology mapping is acceptable should be determined by individual systems (case-by-case). However it's not realistic and possible to include each individual system's mapping requirement and each mapping's accuracy in our massive network analysis. Instead in this paper we evaluate the maximum coverage of semantic interoperability that an information system network could have by accepting partial mappings and information loss. In practice some mapping quality control mechanisms should be introduced for information systems. Mena et al.¹¹ proposed some approaches to estimate the loss of information when a query is translated across different ontologies. Kashyap and Sheth⁶ defined a concept of semantic proximity and further used this concept to determine qualitative measures of semantic similarity between database objects. We believe that similar approaches could be taken to determine whether a mapping is acceptable.

4. Information Fluidity

If two systems are semantically interoperable in the information system network, we say that information can flow from one system 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.

As introduced in Section 3, in the information system network, two semantically interoperable nodes (information systems) are connected with an edge. If any pair of information systems in the network is semantically interoperable, then this network is fully connected. If the network is not fully connected (some systems are not interoperable and isolated in the network), it must consist of several or many fully connected components (sub-networks). Information systems are semantically interoperable within each individual components but not beyond the boundary of components. As shown in Figure 2, in fact information flow is confined within the boundary of these components. Here we propose a new metric called information fluidity to quantify the semantic interoperability level of an information system network. Assume that an information system network includes n systems. With regard to semantic interoperability, the network can be partitioned into k ($k \geq 1$) components and each component includes m_i ($i = 1, 2, \dots, k$) systems. Straightforwardly we have $n = \sum_{i=1}^k m_i$. Define the number of systems in the largest component as m_{max} , i.e., $m_{max} = \max m_i$. Here we use the ratio $\frac{m_{max}}{n}$ to represent the information fluidity of the information system network. In other words, we measure the maximum coverage of systems that information flow can reach. Though many components could exist in a system network, here we only use the size of the largest component to compute the information fluidity because it's hard to calculate the average size of all components in a massive network. As the ratio $\frac{m_{max}}{n}$ increases, the network is claimed to have better information fluidity. For example, every cluster of systems that commit to the same ontology has 100% information fluidity within that cluster since any two systems are semantically interoperable. Straightforwardly, ontology mappings can dramatically improve the information fluidity in a heterogeneous Web since it connects different clusters in the information system network. Note that information fluidity is used to measure the interoperability level of the information system network but not the ontology network.

For convenience in the following sections, here we prove several simple propositions:

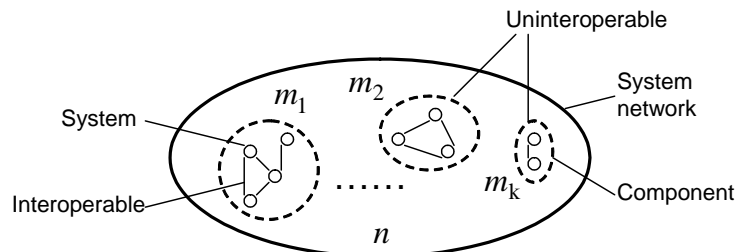


Fig. 2. System network and its components

8 *G. Jiang, G. Cybenko & J. Hendler*

Lemma 4.1. *If two nodes in the ontology network are connected, the associated two clusters of information systems in the system network are semantically interoperable.*

Proof. If two nodes in the ontology network are connected, there exists at least one path (with consecutive edges) that links these two nodes. 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 information systems in the associated two clusters is semantically interoperable via this sequence of ontology mappings. \square

Here we assume that any length of mapping sequence between two nodes is acceptable because certain amount of information (more or less) can flow from one node to another via this path. In practice, a long sequence of ontology mappings may lead to large information loss and computing overhead in data mapping and 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.

Lemma 4.2. *If the ontology network is fully connected, every pair of information systems in the system network is semantically interoperable and the whole information system network has 100% information fluidity.*

Proof. If the ontology network is fully connected, every pair of ontologies in the ontology network is connected. According to Lemma 4.1, every pair of associated clusters of information systems in the system network is semantically interoperable via a sequence of ontology mappings. That means every pair of systems from any two clusters is semantically interoperable via a sequence of ontology mappings. Since all systems within the same cluster commit to the same ontology, they are semantically interoperable with each other. Therefore, every pair of systems in the system network is semantically interoperable with or without ontology mappings. The whole information system network is fully connected with regard to semantic interoperability. According to our definition of information fluidity, since 100% systems in the information system network are semantically interoperable, the whole network has 100% information fluidity. \square

5. Organized Ontology Network

According to Lemma 4.2, if the ontology network 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?

Lemma 5.1. *In an ontology network with M ontologies, at least $M - 1$ ontology mappings and at most $\frac{(M-1)(M-2)}{2} + 1$ 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). For example, link these M vertices with a line. Conversely, in the worst case, it could take $\frac{(M-1)(M-2)}{2} + 1$ edges to get a graph with M vertices fully connected, i.e., $M - 1$ vertices are fully meshed before a new edge is added to get the last isolated vertex connected. According to Lemma 4.2, with a fully connected ontology graph, the associated information system network has 100% information fluidity. \square

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 information 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 Lemma 5.1, 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 Network

In a small network, 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.⁴ 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.^{12,13} Barabasi et al.¹⁴ 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 Lu¹⁵ 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.,

$$p_{ij} = \frac{w_i w_j}{\sum_{k=1}^n w_k}, \text{ assuming } \max w_i^2 \leq \sum_{k=1}^n w_k. \quad (6.1)$$

This assumption ensures that $p_{ij} \leq 1$ for all i and j . According to Eq. (6.1), 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 $d = \sum_{i=1}^n w_i/n$. 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 Lu¹⁵ proved the following theorem. Here “almost surely” means that the following results hold with probability one.

Theorem 6.1. (Chung and Lu 2002) For a random graph $G(w)$ with a given expected degree sequence having average degree $d > 1 + \delta > 1$, almost surely $G(w)$ has a unique giant component.

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

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

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

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

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 discussed in Section 3, every information system 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. Assume that each cluster includes n_i ($i = 1, 2, \dots, M$) systems and thus $\sum_{i=1}^M n_i = N$. Define $n_{max} = \max n_i$, i.e., n_{max} is the size of the largest system cluster that commits to the same ontology.

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 number of its instances in the system network. We define

$$w_i = K \cdot \frac{n_i}{N}, \quad i = 1, 2, \dots, M, \quad (6.2)$$

where K is the sum of all nodes’ degrees in the ontology graph. In fact, with Eq. (6.2), we have

$$\sum_{i=1}^M w_i = \sum_{i=1}^M K \cdot \frac{n_i}{N} = \frac{K}{N} \cdot \sum_{i=1}^M n_i = K. \quad (6.3)$$

Meantime, as shown in Eq. (6.1), 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.,

$$p_{ij} = \frac{w_i w_j}{\sum_{k=1}^M w_k} = K \cdot \frac{n_i n_j}{N^2} \quad \text{with } K \leq \frac{N^2}{n_{max}^2}. \quad (6.4)$$

What does Eq. (6.4) mean in our model? Building semantic mapping between two ontologies is likely a bilateral activity and both sides could be motivated to map terminologies and structures between ontologies to achieve better information fluidity. 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 build 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 $d = \sum_{i=1}^M w_i / M = \frac{K}{M}$ and according to

12 *G. Jiang, G. Cybenko & J. Hendler*

Eq. (6.2), the volume of subset S is

$$\text{Vol}(S) = \sum_{v_i \in S} w_i = \frac{K}{N} \cdot \sum_{v_i \in S} n_i. \quad (6.5)$$

The volume of the whole ontology network G is

$$\text{Vol}(G) = \sum_{i=1}^M w_i = K. \quad (6.6)$$

6.3. Lower bound of information fluidity

With the above model and parameters, given an information system network, we can derive from Theorem 6.1 the following lower bound of its information fluidity.

Theorem 6.2. 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 expected edge number of an ontology node 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 total 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 constraint in Eq. (6.4), $L = \frac{K}{2} \leq \frac{N^2}{2n_{max}^2}$, 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 6.1, 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

$$\text{Vol}(S_{giant}) \geq (1 - \frac{2}{\sqrt{d \cdot e}} + o(1)) \text{Vol}(G). \quad (6.7)$$

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

$$\frac{\text{Vol}(S_{giant})}{\text{Vol}(G)} \geq 1 - \frac{2}{\sqrt{d \cdot e}} + o(1) = 1 - \sqrt{\frac{2M}{L \cdot e}} + o(1). \quad (6.8)$$

Meantime, by the definition of $\text{Vol}(S)$ and $\text{Vol}(G)$ in Eq. (6.5) and Eq. (6.6) respectively, we have

$$\frac{\text{Vol}(S_{giant})}{\text{Vol}(G)} = \frac{\frac{K}{N} \cdot \sum_{v_i \in S_{giant}} n_i}{K} = \frac{\sum_{v_i \in S_{giant}} n_i}{N} \quad (6.9)$$

Therefore by Inequality (6.8) and Eq. (6.9), we have

$$\frac{\sum_{v_i \in S_{giant}} n_i}{N} \geq 1 - \sqrt{\frac{2M}{L \cdot e}} + o(1). \quad (6.10)$$

The giant component S_{giant} is the largest sub-graph that is fully connected in the ontology network. For any node (ontology) $v_i \in S_{giant}$ in the ontology network, n_i systems in the information system network commit to that ontology. Therefore, a total of $\sum_{v_i \in S_{giant}} n_i$ systems in the information 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 Lemma 4.2, any pair of these $\sum_{v_i \in S_{giant}} n_i$ systems is semantically interoperable with or without ontology mappings. Therefore in the information system network, the largest percentage of systems that are semantically interoperable between each other is at least $\sum_{v_i \in S_{giant}} n_i / N$. By the metric definition of information fluidity and Inequality (6.10), 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 6.1. \square

6.4. Result analysis

In this section, we use an example to illustrate Theorem 6.2. Assume that the ontology network has $M = 5000$ ontologies and the largest cluster in the information

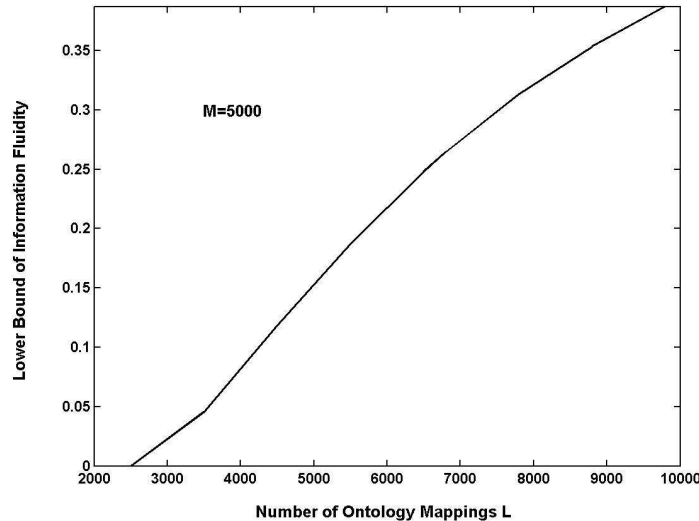


Fig. 3. Information fluidity vs. number of ontology mappings ($M = 5000$)

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 constraint in Eq. (6.4), we have $L \leq \frac{N^2}{2m_{max}^2} = 20000$. 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 Eq. (6.4). By the first result of Theorem 6.2, 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 3 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 such a large ontology network. Given different sizes of ontology networks, Figure 4 shows the growth of information fluidity as the number of ontology mappings increases. Given different number of ontology mappings, Figure 5 illustrates how the information fluidity decreases as the size of ontology networks grows. Note that all these curves

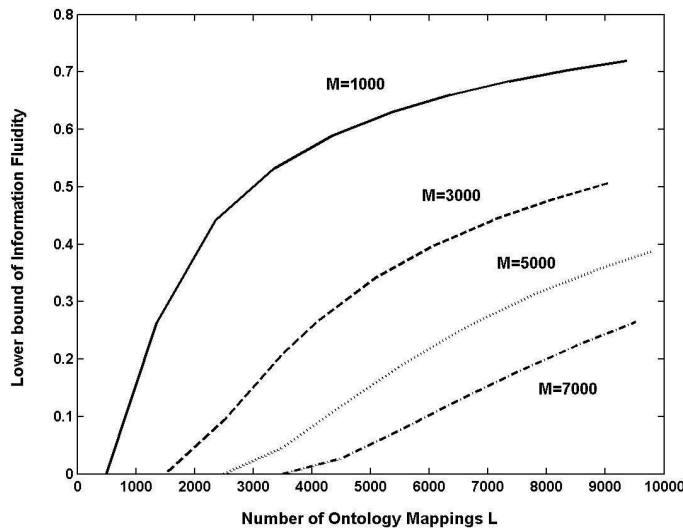


Fig. 4. Information fluidity growth with different sizes of ontology networks

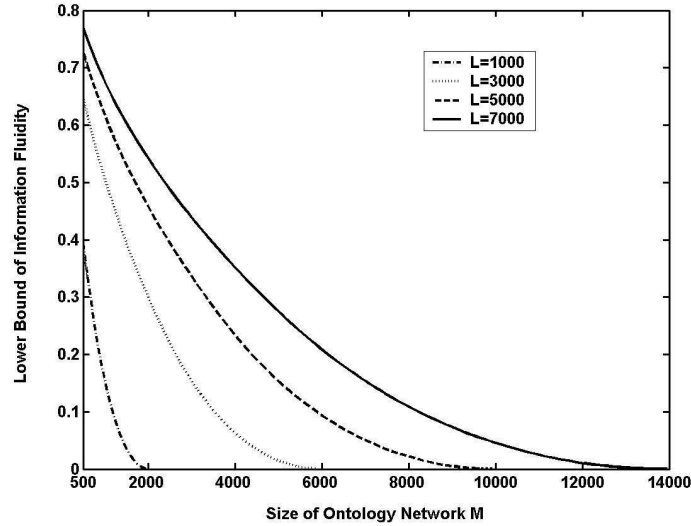


Fig. 5. Information fluidity vs. size of ontology networks

only reflect the lower bound of information fluidity and the real interoperability level of an information system network is always higher than this bound.

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 4, 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 Eq. (6.4), which leads to the faster growth of ontology network connectivity and further the faster growth of information fluidity in the system network.

In Figure 4, 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 Eq. (6.4). However, as we see in Figure 3 and 4, 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: $\frac{(M-1)(M-2)}{2} + 1$. For $M = 1000$, this number is as high as half million.

6.5. Applications

As discussed in Section 4, we develop the concept of information fluidity to measure the semantic interoperability level of a large information system network. This metric captures the mass characteristics of system networks and enables us to have a macroscopic view on their semantic interoperability. Many details about the heterogeneity of systems, ontologies and mappings are abstracted so as to develop a single measurement to evaluate the system interoperability level of a massive network. In practice, we believe that it’s necessary to ignore such details and select an appropriate abstract level to characterize a large system network. An analogy is that the concepts of “Mean” and “Variance” are used to describe the property of a data set though they do not capture the data distribution precisely.

After we have this metric to evaluate the semantic interoperability level, our results can be useful in evaluating mapping efforts needed for large-scale heterogeneous information systems. Given the numbers of ontologies and systems, based on Theorem 6.2, we can calculate the ontology mapping efforts needed to achieve a targeted level of semantic interoperability. Conversely, given the numbers of ontologies and their mappings, we can also calculate the semantic interoperability level that could result from these mappings. Therefore the definition of information fluidity and our analytical results enable us to do cost benefit analysis in engineering. For example, today’s battlefield information systems such as C4ISR consist of thousands of sensors, weapons, platforms and intelligence databases etc. These systems employ hundreds of various schemas/ontologies to describe their data. Without ontology mappings, information flow is blocked at the boundary of these ontologies. Therefore one critical challenge is how to improve the semantic interoperability among these systems in order to support dynamic information sharing in battlefields. For such a large scale information system network, before starting ontology mapping efforts, it’s very necessary to evaluate the potential semantic interoperability level that could benefit from the resource invested for such mapping efforts. If the semantic interoperability is improved little after a large number of ontology mappings, it’s probably better to choose other approaches such as deploying common ontologies mandatorily. As shown in Figure 4 and 5, our analytical results enable us to roughly estimate the interoperability level of such a large scale information system network and support cost benefit analysis.

7. Discussions

For a market-driven ontology network, Theorem 6.2 gives the lower bound of information fluidity in its associated information system network, i.e. the worst-case scenario. The real interoperability level of the information system network will be higher though it's not clear how to measure this metric instead of the lower bound. Meantime, many fully connected components may exist in the system network. All information systems within the same component are semantically interoperable and each component has 100% information fluidity within its boundary. However, as shown in Figure 2, these components are not connected in the information system network. Without ontology mapping, information flow is blocked at the boundary of system clusters that commit to the same ontology. With ontology mapping, it is blocked at the boundary of these components. According to our metric definition for information fluidity, we only use the size of the largest component in a system network to quantify the information fluidity because it's hard to calculate the average size of components in a massive network. Therefore, our metric doesn't indicate that only that percentage of systems is semantically interoperable. Instead, many systems may have semantic interoperability within their smaller component but not beyond the boundary of that component.

According to Eq. (6.1), 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 ontology nodes in the ontology graph. Though this limitation could make our lower bound in Theorem 6.2 not tight, all conclusions in Theorem 6.2 will still be valid. We will address this limitation of our model in our future work. 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, as illustrated in Section 6, this number is still large enough for us to analyze different level of information fluidity with growing mappings. Meantime, we argue that the ontology network 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. In fact, we are specifically interested in the information fluidity growth of this phase transition stage. Once after the ontology network is fully connected, the information fluidity of the system network will always be 100% and it's not necessary to analyze the component size anymore. As discussed in Section 3, one ontology mapping could consist of several partial mappings, i.e., one ontology can be split into several parts and then each part is mapped into the target ontology via different mapping paths in the network. Since we consider two ontologies are connected in the ontology graph as long as there is one path between them, the multiple paths (multiple partial mappings) between two ontologies will not affect our measure of information fluidity though such multiple mappings could reduce information loss during communication.

Note that the graph model $G(w)$ described in Section 6.1 is a general random graph model with given expected degree sequences. The common Erdos-Renyi model¹⁶ $G(n, p)$ can be viewed as a special case of $G(w)$ with all w_i (degrees of nodes) equal. In this case, the probability p to connect any two nodes in the graph is equal according to Eq. (6.1). Conversely, in our model, the degree of an ontology node is proportional to the number of its instances in the system network. Two “popular” nodes are likely to be connected with a high probability p as shown in Eq. (6.4). If we choose the degree sequence $w = (w_1, w_2, \dots, w_n)$ satisfying $w_i = c \cdot i^{-1/(\beta-1)}$ for $i_0 \leq i \leq n + i_0$. Here c is determined by the average degree d and i_0 depends on the maximum degree m , namely, $c = \frac{\beta-2}{\beta-1} d n^{-1/(\beta-1)}$ and $i_0 = n \left(\frac{d(\beta-2)}{m(\beta-1)} \right)^{\beta-1}$. According to Ref. 17, then the degree distribution of $G(w)$ follows power-law distribution, i.e., the density function $p(k) \leftarrow k^{-\beta}$. Since the power-law distribution can be transformed to a degree distribution of the graph model $G(w)$ in Section 6.1, straightforwardly both Theorem 6.1 and Theorem 6.2 will still hold for this special case. For power-law distribution network, there are literatures to roughly estimate the order of component size under different β . For example, for $1 < \beta < 2$, it was proved that the network has a giant component of size $\Theta(n)$.¹⁸ However, to the best of our knowledge, we have not seen any so accurate bound estimation as shown in Theorem 6.1 and 6.2.

As discussed earlier, many details about the heterogeneity of systems, ontologies and mappings are abstracted so as to develop a single measurement of information fluidity. In our future work, we plan to analyze semantic interoperability at a finer granularity so as to estimate information fluidity more accurately. We also need to consider how to incorporate mapping quality and information loss in the information fluidity analysis. Instead of analyzing interoperability at information system level, one alternative is to analyze interoperability at the level of concepts embedded in ontologies. For example, if we replace the ontology nodes in Figure 1 with single concepts, we are able to get similar information fluidity measurement with respect to concepts. However, since system interoperability is usually analyzed among individual systems and one system may employ many concepts from its committed ontology, it’s unclear how to aggregate the interoperability at the concept level to obtain the interoperability measurement at the system level. We will attack these problems in our future work.

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.¹⁹ The power-law distribution has been applied to model the connectivity of nodes in many complex networks. Recently, Newman²⁰ and Albert et al.¹³ 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 hyper-link structure of the World Wide Web. Barabasi and Albert²¹ and Broder et al.²² 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.²³, Albert and Barabasi¹³ and Aiello et al.¹⁸ 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.²⁴ 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 existing 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^{2,4} and in analyzing the advantage of an approach with many ontologies mapped to each other.

9. Conclusions

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 formulated the information fluidity problem as a graph connectivity problem. For market-driven ap-

proach 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. We believe that our model and results can be useful in evaluating mapping efforts needed for large-scale heterogeneous information systems.

Acknowledgments

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. The work was performed when the first author worked at the Institute of Security Technology Studies at Dartmouth College. We also want to thank anonymous reviewers for their valuable comments, especially the reviewer who encouraged us to analyze the interoperability at the concept level.

References

1. World Wide Consortium (W3C): Semantic Web Activity. [Http://www.w3.org/2001/sw/](http://www.w3.org/2001/sw/).
2. T. Berners-Lee, J. Hendler and O. Lassila, The semantic web, *Scientific American*. **284**(5) (2001) 34-43.
3. T. Gruber, A translation approach to portable ontologies, *Knowledge Acquisition*. **5**(2) (1993) 199-220.
4. J. Hendler, Agents on the semantic web, *IEEE Intelligent Systems*. **16**(2) (2001) 30-37.
5. H. Chalupsky, OntoMorph: a translation system for symbolic knowledge, in *Proc. of the Seventh International Conference on Principles of Knowledge Representation and Reasoning*, eds. A. Cohn, F. Giunchiglia and B. Selman (2000) pp. 471-482.
6. V. Kashyap and A. Sheth, Semantic and schematic similarities between database objects: a context approach, *VLDB Journal*. **5**(4) (1996) 276-304.
7. W. Kim and J. Seo, Classifying schematic and data heterogeneity in multidatabase systems, *IEEE Computer*. **24**(12) (1991) 12-18.
8. D. Dou, D. McDermott and P. Qi, Ontology translation by ontology merging and automated reasoning, in *Proc. of EKAW Workshop on Ontologies for Multi-Agent Systems*, (2002).
9. A. Doan, J. Madhavan, P. Domingos and A. Halevy, Learning to map between ontologies on the semantic web, in *Proc. of World Wide Web Conference 2002*, (2002) pp. 662-673.
10. E. Mena, A. Illarramendi, V. Kashyap and A. Sheth, OBSERVER: an approach for query processing in global information systems based on interoperation across pre-existing ontologies, *International Journal on Distributed And Parallel Databases*. **8**(2) (2000) 223-271.

11. E. Mena, V. Kashyap, A. Illarramendi and A. Sheth, Imprecise answers on highly open and distributed environments: an approach based on information loss for multi-ontology based query processing, *International Journal of Cooperative Information Systems (IJCIS)*. **9**(4) (2000) 403-425.
12. W. Aiello, F. Chung and L. Lu, Random evolution in massive graphs, *Handbook on Massive Data Sets*(Kluwer Academic Publishers, 2002), eds. James Abello et al., pp. 97-122.
13. R. Albert and A. Barabasi, Statistical mechanics of complex networks, *Reviews of Modern Physics*. **74**(1) (2002) 47-97.
14. A. Barabasi, R. Albert and H. Jeong, Mean-field theory for scale-free networks, *Physica A*. **272** (1999) 173-187.
15. F. Chung and L. Lu, Connected components in random graphs with given expected degree sequences, *Annals of Combinatorics*. **6**(2) (2002) 125-145.
16. B. Bollobas, *Modern Graph Theory* (Springer, 1998).
17. F. Chung, L. Lu and V. Vu, Spectra of random graphs with given expected degrees, *Proc. of National Academy of Science*. **100**(11) (2003) 6313-6318.
18. W. Aiello, F. Chung and L. Lu, A random graph model for massive graphs, in *Proc. of the Thirty-second Annual ACM Symposium on Theory of Computing*, (2000) 171-180.
19. D. Watts and S. Strogatz, Collection dynamics of "small-world" networks, *Nature*. **393** (1998) 440-442.
20. M. Newman, The structure and functions of complex networks, *SIAM Review*. **45**(2) (2003) 167-256.
21. A. Barabasi and R. Albert, Emergence of scaling in random networks, *Science*. **286**(509) (1999) 509-512.
22. A. Broder, R. Kumar, F. Maghoul, S. Rajagopalan, R. Stata, A. Tomkins and J. Wiener, Graph structure in the web, in *Proc. of the 9th WWW/Computer Networks 33*, (2000) pp. 309-320.
23. R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins and E. Upfal, Stochastic models for the web graph, in *Proc. of the 41st IEEE Conference on Foundation of Computer Science*, (2000) pp. 57-66.
24. S. Dill, R. Kumar, K. Mccurley, S. Rajagopalan, D. Sivakumar and A. Tomkins, Self-similarity in the web, *ACM Transactions on Internet Technology*. **2**(3) (2002) 205-223.