Analyzing Semantic Interoperability in Bioinformatic Database Networks[*]

Philippe Cudré-Mauroux, Julien Gaugaz, Adriana Budura, and Karl Aberer

School of Computer and Communication Sciences
EPFL – Switzerland

Abstract. We consider the problem of analytically evaluating semantic interoperability in large-scale networks of schemas interconnected through pairwise schema mappings. Our heuristics are based on a graph-theoretic framework capturing important statistical properties of the graphs. We validate our heuristics on a real collection of interconnected bioinformatic databases registered with the Sequence Retrieval System (SRS). Furthermore, we derive and provide experimental evaluations of query propagation on weighted semantic networks, where weights model the quality of the various schema mappings in the network.

1 Introduction

Even if Semantic Web technologies have recently gained momentum, their deployment on the wide-scale Internet is still in its infancy. Only a very small portion of websites have so far been enriched with machine-processable data encoded in RDF or OWL. Thus, the difficulty to analyze semantic networks due to the very lack of realistic data one can gather about them. In [5], we introduced a graph-theoretic model to analyze interoperability of semantic networks and tested our heuristics on large-scale, random topologies. In this paper, we extend these heuristics and test them on a real semantic network, namely on a collection of schemas registered with the Sequence Retrieval System (SRS).

We start below by giving a short introduction to SRS. We then present our approach, which boils down to an analysis of the component sizes in a graph of schemas interconnected through schema mappings. We report on the statistical properties of the SRS network we consider and on the performance of our heuristics applied on this network. Finally, we report on the performance of our approach on larger and weighted networks mimicking the statistical properties of the SRS network.

[*] The work presented in this paper was supported (in part) by the National Competence Center in Research on Mobile Information and Communication Systems (NCCR-MICS), a center supported by the Swiss National Science Foundation under grant number 5005-67322.

2 The Sequence Retrieval System (SRS)

SRS, short for "Sequence Retrieval System", is a commercial information indexing and retrieval system designed for bioinformatic libraries such as the EMBL nucleotide sequence databank, the SwissProt protein sequence databank or the Prosite library of protein subsequence consensus patterns. It is a distributed, interoperable data management system which was initially developed at the European Molecular Biology Laboratory in Heidelberg, and which allows the querying of one or several databases simultaneously, regardless of their format or schemas.

Administrators wishing to register new databases with SRS first have to define the schema they have adopted to store data, using a custom language called *Icarus*. Once their schemas have been defined, administrators can import schema instances (i.e., text files) whose data will be correctly parsed, indexed and processed thanks to the corresponding schema definitions. Additionally, administrators can manually define relationships between their database schema and similar schemas. In SRS, these relationships are represented as links relating one entry of a database schema to one entry of another schema. Thanks to this structure of links between databases, users can propagate queries they pose locally against one specific schema to other schemas available in the system (for technical details, we refer the interested reader to *http://www.lionbioscience.com/*)

2.1 Graph analysis of an SRS repository

Conceptually, the model described above is very close to what one could expect from a subgraph of the semantic web itself, i.e., a collection of related schemas (or ontologies) linked one to another through pairwise mappings. The graph which can be extracted from a SRS repository has two main advantages over those which can be built from current RDFS / OWL repositories: i) it is based on a real-world collection of schemas which are being used on a daily basis by numerous independent parties and ii) it is of a reasonable size (several hundreds of semantically related complex schemas). Thus, after having been rather unsuccessful at finding reasonable semantic networks from the Semantic Web itself, we decided to build a specialized crawler to analyze the semantic graph of an SRS repository and to test our heuristics on the resulting network.

We chose to analyze the semantic network from the European Bioinformatics Institute SRS repository, publicly available at *http://srs.ebi.ac.uk/*. We built a custom crawler which traverses the entire network of databases and extracts schema mapping links stored in the schema definition files. The discussion below is based on the state of the SRS repository as of May 2005.

The graph resulting from our crawling process is an undirected graph of 388 nodes (database schemas) and 518 edges (pairwise schema mappings). We chose to represent links as undirected edges since they are used in both directions by SRS (they basically represent cross-references between two entries of two database schemas). We identified all connected components in the graph (two nodes are in the same connected component if there is a path from one to the other following edges). The analysis revealed one giant connected component (i.e., a relatively large set of interconnected schemas) of 187 nodes, which represent roughly half of the nodes and 498 edges. Besides the giant connected component, the graph also has two smaller components, each consisting of two vertices. The rest of the nodes are isolated, representing mostly result databases or databases for which no link to other databases was defined.

The average degree of the nodes is 2.2 for the whole graph and 4.6 for the giant component. Real networks differ from random graphs in that often their degree distribution follows a power law, or has a power law tail, while random graphs have a Poisson distribution of degrees [2]. Unsurprisingly, our semantic network is no exception as can be seen in Figure 1 below, which depicts an accurate approximation of the degree distribution of our network by a power-law distribution $y(x) = \alpha x^{-\gamma}$ with $\alpha = 0.21$ and $\gamma = 1.51$.

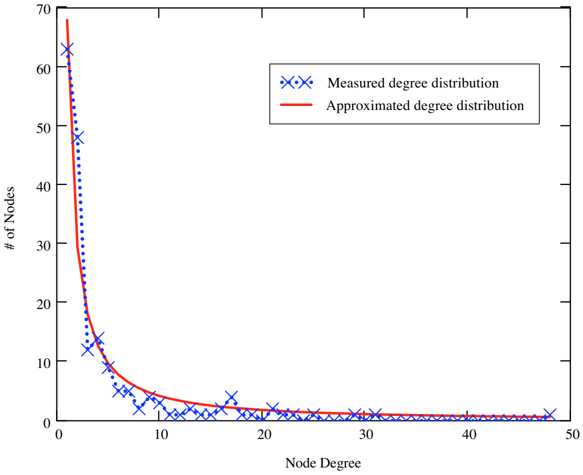

Figure 1: An approximation of the degree distribution of our semantic network by a power-law distribution $y(x) = \alpha x^{-\gamma}$ with $\alpha = 0.21$ and $\gamma = 1.51$

Another interesting property which we explored was the tendency of the schemas to form clusters, quantified by the average clustering coefficient. Intuitively, the clustering coefficient of a vertex measures the degree to which its neighbors are neighbors of each other. More precisely, the clustering coefficient indicates the ratio of existing edges connecting a node's neighbors to each other to the maximum possible number of such edges. The network we considered has a high average clustering coefficient of 0.32 for the whole graph and of 0.54 for the giant component. The diameter (maximum shortest paths between any two vertices) of the giant connected component is 9. These data indicate that our network can be characterized as *scale-free* (power-law distribution of degrees) or *small-world* (small diameter, high clustering coefficient).

3 Analyzing semantic interoperability in the large

In [5], we introduced a graph-theoretic framework for analyzing semantic interoperability in large-scale networks. As described above, we model database schemas (or ontologies) as nodes, interconnected by edges (schema mappings). Schema mappings are used to iteratively propagate queries posed against a local schema to other related schemas (see [1] for how this can be implemented in practice). Note that links can be directed or undirected, weighted or non-weighted depending on the schema mappings being used.

In such a network, the density of mappings is important in order to propagate a local query from one database (schema) to the other databases. A query can only be propagated to all databases if the semantic network is *connected*, that is if there exists a path from one schema to any other schema following schema mapping links. If some schemas are isolated, queries cannot be propagated to/from the rest of the graph, thus making it impossible to have a semantically interoperable network of databases. This observation motivated us to take advantage of percolation theory to determine when a semantic network could be connected or not. Our framework for analyzing semantic interoperability takes advantage of *generatingfunctionologic* [7] functions for the degree distribution of the semantic graph:

$$G_0(x) = \sum_{k=0}^{\infty} p_k x^k \tag{1}$$

where p_k represents the probability that a randomly chosen vertex has degree k. We showed (by extending results from [6]) that a network cannot be semantically interoperable in the large unless the *connectivity indicator* $ci = \sum_k k(k-2-cc)p_k$ is greater than zero, with cc representing the clustering coefficient. Also, we provided heuristics for estimating the relative size S of the biggest semantically interoperable cluster of schemas by solving

$$S = 1 - G_0(u), \tag{2}$$

where u is the smallest non-negative real solution of

$$u = G_1(u) \tag{3}$$

and $G_1(u)$, the distribution of outgoing edges from first to second-order neighbors, is

$$G_1(x) = \frac{1}{x^{cc}} \frac{G_0'(x)}{G_0'(1)} = \frac{1}{z_1} \frac{1}{x^{cc}} G_0'(x). \tag{4}$$

3.1 Applying our heuristics to the SRS graph

We applied our heuristics to the SRS graph we obtained from the crawling process. The results are as follows: we get a connectivity indicator ci of 25.4, indicating that the semantic network is clearly in a super-critical state and that a giant component interlinking most of the databases has appeared. The size of this giant component as estimated by our heuristics (see above) is of 0.47, meaning that 47% of the schemas are part of the giant connected component. This is surprisingly close to the real value of 0.48 as observed in the graph.

3.2 Generating a Graph with a given Power-Law Degree Distribution

Going slightly further, we want to analyze the dynamics of semantic graphs with varying numbers of edges. Our aim is to generate graphs with the same statistical properties as the SRS graphs, that is, graphs following a power-law degree distribution:

$$P(k) = \alpha k^{-\gamma} \tag{5}$$

but with a varying number of edges. We take from [3] a graph-building algorithm yielding a power-law degree distribution with a given exponent γ. It goes as following: (1) create a (large) number N of vertices. (2) to each vertex i, assign an "importance" x_i, which is a real number taken from a distribution $\rho(x)$. (3) for each pair of vertices, draw a link with probability $f(x_i, x_j)$, which is function of the importance of both vertices.

Now if $f(x_i, x_j) = (x_i x_j / x_M^2)$ (where x_M is the largest value of x in the graph), we know from [3] that the degree distribution of a graph will be

$$P(k) = \frac{x_M^2}{N\langle x\rangle}\rho\left(\frac{x_M^2}{N\langle x\rangle}k\right) \tag{6}$$

where $\langle x\rangle$ is the expected value of the importance x, such that $P(k)$ follows a power-law if $\rho(x)$ does so.

We then choose a power-law distribution

$$\rho(x) = \frac{\gamma - 1}{(m^{1-\gamma} - Q^{1-\gamma})} x^{-\gamma} \tag{7}$$

defined over the interval $[m, Q]$. However, we still have to find values for m and Q such that the scale of the resulting degree distribution equals α. Using equation 7, we find the expected importance value as

$$\langle x\rangle = \frac{(\gamma - 1)(m^2 Q^\gamma - m^\gamma Q^2)}{(\gamma - 2)(m Q^\gamma - m^\gamma Q)}. \tag{8}$$

Replacing $\rho(x)$ in equation 6, the degree distribution of the resulting graph becomes

$$P(k) = \frac{x_M^2}{N\langle x\rangle}\frac{\gamma - 1}{(m^{1-\gamma} - Q^{1-\gamma})}\left(\frac{x_M^2}{N\langle x\rangle}k\right)^{-\gamma} \tag{9}$$

such that, equating with equation 5, we get

$$\alpha = \frac{x_M^2}{N\langle x\rangle} \frac{\gamma - 1}{(m^{1-\gamma} - Q^{1-\gamma})} \left(\frac{x_M^2}{N\langle x\rangle} \right)^{-\gamma}. \tag{10}$$

We can then arbitrarily choose $m > 0$ and find Q by numerical approximation, since the right-hand side of equation 10 is defined and continue for values of $Q > m$.

Figures 2 and 3 show the results of our heuristics on networks of respectively 388 (i.e., mimicking the original SRS graph) and 3880 edges (i.e., 10 times bigger than the original SRS graph but with the same statistical properties) constructed in the way presented above with a varying number of edges. The curves are averaged over 50 consecutive runs. As for the original SRS network, we see that we can accurately predict the size of the giant semantic component, even for very dense graphs.

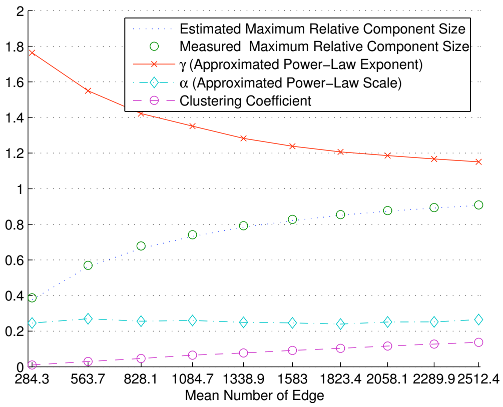

Figure 2: Estimating the giant component size of a scale-free semantic network of 388 nodes with a varying number of edges

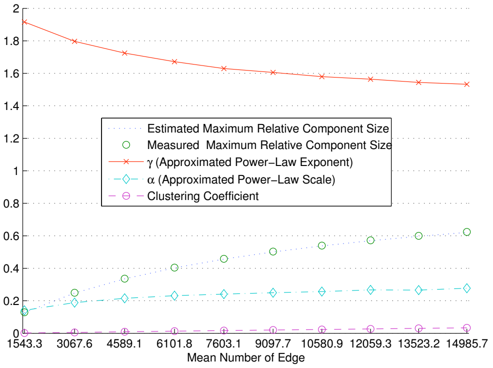

Figure 3: Estimating the giant component size of a scale-free semantic network of 3880 nodes with a varying number of edges

4 Connectivity Indicator and Giant Component Size in Weighted Graphs

So far, we only analyzed the presence and size of a giant connected component in order to determine which portion of a semantic network could potentially be semantically interoperable. In large-scale decentralized networks, however, one should not only look into the giant semantic component itself, but also analyze the *quality* of the mappings used to propagate queries from one schema to the other (see [1] for a discussion on that topic). Indeed, in any large, decentralized network, it is very unlikely that all schema mappings could *correctly* map queries from one schema to the other, because of the lack of expressivity of the mapping languages, and of the fact that some (most?) of the mappings might be generated automatically.

Thus, as considered by more and more semantic query routing algorithms, we introduce weights for the schema mappings to capture the quality of a given mapping. Weights range from zero (indicating a really poor mapping unable to semantically keep any information while translating the query) to one (for perfect mappings, keeping the semantics of the query intact from one schema to the other). We then iteratively forward a query posed against a specific schema to other schemas through schema mappings if and only if a given mapping has a weight (i.e., quality) greater than a predefined threshold τ. $\tau = 0$ corresponds to sending the query through any schema mapping, irrespective of its quality. On the contrary, when we set τ to one, the query gets only propagated to semantically perfect mappings, while even slightly faulty mappings are ignored. Previous works in statistical physics and graph theory have looked into percolation for weighted graphs. We present hereafter an extension of our heuristics for weighted semantic networks inspired by [4].

4.1 Connectivity Indicator

As before, we consider a generating function for the degree distribution

$$G_0(x) = \sum_{k=0}^{\infty} p_k x^k \tag{11}$$

where p_k is the probability that a randomly chosen vertex has degree k in the network. We then define t_{jk} as the probability that an edge has a weight above τ given that it binds vertices of degree j and k. Thus, $w_k = \sum_{j=0}^{\infty} t_{jk}$ is the probability that an edge transmits, given that it is attached to a vertex of degree k. The generating function for the probability that a vertex we arrive at by following a randomly chosen edge is of degree k *and* transmits is

$$G_1(x) = \frac{\sum_{k=0}^{\infty} w_k k x^{k-1}}{x^{cc} \sum_{k=0}^{\infty} k p_k} \tag{12}$$

where cc is the clustering coefficient. Now, from [4], we know that the generating function for the probability that one end of a randomly chosen *edge* on the graph leads to a percolation cluster of a given number of vertices is

$$H_1(x) = 1 - G_1(1) + xG_1\left[H_1(x)\right]. \tag{13}$$

Similarly, the generating function for the probability that a randomly chosen *vertex* belongs to a percolation cluster of a given number of vertices is

$$H_0(x) = 1 - G_0 + xG_0\left[H_1(x)\right] \tag{14}$$

such that the mean component size corresponding to a randomly chosen vertex is

$$\langle s \rangle = H_0'(1) = G_0(1) + \frac{G_0'(1)G_1(1)}{1 - G_1'(1)} \tag{15}$$

which diverges for $G_1'(1) \geq 1$. However,

$$G_1'(1) = \frac{\sum_{k=0}^{\infty} w_k p_k k(k-1-cc)}{\sum_{k=0}^{\infty} kp_k} \tag{16}$$

such that a giant connected component appears if

$$ci = \sum_{k=0}^{\infty} kp_k\left[w_k(k-1-cc) - 1\right] \geq 0 \tag{17}$$

4.2 Giant Component Size

As seen above, $H_0(x)$ represents the distribution for the cluster size which a randomly chosen vertex belongs to, *excluding* the giant component. Thus, according to [4], $H_0(1)$ is equal to the fraction of the nodes which are not in the giant component. The fraction of the nodes which are in the giant component is hence $S = 1 - H_0(1)$. Using equation 14 we can write

$$S = 1 - H_0(1) = G_0(1) - G_0\left[H_1(1)\right]. \tag{18}$$

with $H_1(1) = 1 - G_1(1) + xG_1\left[H_1(1)\right]$. Thus $H_1(1) = u$ where u is the smallest non-negative solution of

$$u = 1 - G_1(1) + G_1(u). \tag{19}$$

The relative size of the giant component reached by the query in a weighted semantic graph follows as

$$S = G_0(1) - G_0(u). \tag{20}$$

Figures 4 and 5 show the results of our heuristics on weighted networks of respectively 388 and 3880 nodes, for a varying number of edges and various values of τ. The curves are averaged over 50 consecutive runs, and the weights of individual schema mappings are randomly generated using a uniform distribution ranging from zero to one. We see that our heuristics can quite adequately predict the relative size of the graph to which a given query will be forwarded. Also, as for the unweighted analysis, we observe similar behaviors for the two graphs; This is rather unsurprising as we are dealing with scale-free networks whose properties are basically independent of their size.

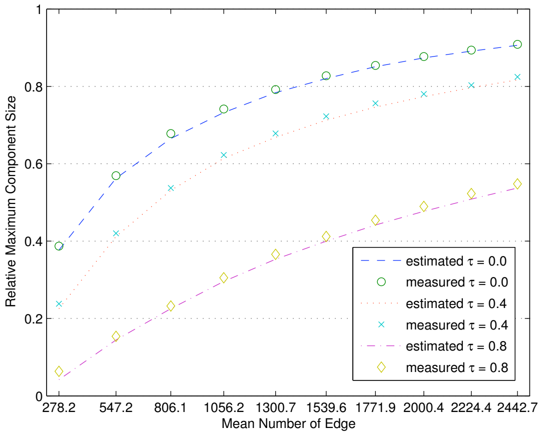

Figure 4: Fraction of the graph a local query will be forwarded to, for a weighted network of 388 nodes with a varying number of edges and various forwarding thresholds τ

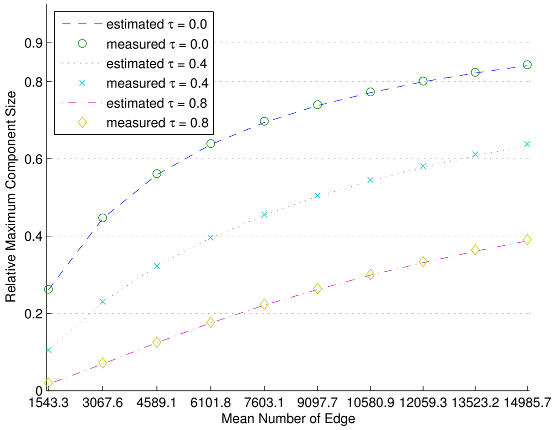

Figure 5: Fraction of the graph a local query will be forwarded to, for a weighted network of 3880 nodes with a varying number of edges and various forwarding thresholds τ

5 Conclusions

In this paper, we tested graph-theoretic heuristics to evaluate semantic interoperability on a real semantic network. The results confirm the validity of our heuristics beyond our initial hopes as we could predict quite accurately the size of the giant semantic component in the network. Also, we extended our analysis to apply our heuristics on larger networks enjoying similar statistical properties and on weighted semantic networks. It was for us quite important to test our heuristics using real-world data, as semantic network analyses mostly consider artificial networks today, due to the current lack of semantically enriched websites or deployed semantic infrastructures. In the future, we plan to extend our analyses to take into account the dynamicity (churn) of the network of schema mappings, and to consider more accurate query forwarding schemes based on transitive closures of mapping operations.

References

1. K. Aberer, P. Cudré-Mauroux, and M. Hauswirth. The Chatty Web: Emergent Semantics Through Gossiping. In *International World Wide Web Conference (WWW)*, 2003.
2. R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(47), 2002.
3. G. Caldarelli, A. Capocci, P. D. L. Rios, and M. Muoz. Scale-free networks from varying vertex intrinsic fitness. *Phys. Rev. Lett.*, 89, 258702, 2002.
4. D. S. Callaway, M. Newmann, S. H. Strogatz, and D. J. Watts. Network robustness and fragility: Percolation on random graphs. *Phys. Rev. Lett.*, 85, 54685471, 2000.
5. P. Cudré-Mauroux and K. Aberer. A Necessary Condition For Semantic Interoperability In The Large. In *International Conference Ontologies, DataBases, and Applications of Semantics (ODBASE)*, 2004.
6. M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev.*, E64(026118), 2001.
7. H. S. Wilf. *Generatingfunctionology.* 2nd Edition, Academic Press, London, 1994.