COSIN

COSIN

IST–2001-33555

COevolution and Self-organization In dynamical Networks

DATABASE OF COLLECTED DATA

Deliverable Number:	D12
Delivery Date:	March 2004
Classification:	Public
Partner owning:	**CR4** (UNIL University of Lausanne)
Contact authors:	Fabrizio Coccetti (C01), Paolo De Los Rios (CR4)
Project Co-ordinator:	Guido Caldarelli (INFM) *guido.caldarelli@roma1.infn.it*, Istituto Nazionale Fisica per la Materia
Partners:	**CR2** (UDRLS) Università "La Sapienza" *Italy* **CR3** (UB) Universitat de Barcelona, *Spain* **CR4** (UNIL), Université de Lausanne *Switzerland* **CR5** (ENS) Ecole Normale Superieure, *France* **CR7** (UNIKARL), University of Karlsruhe, German*y* **CR8** (UPSUD) Université de Paris-Sud, *France*

Information Society

Project funded by the European Community under the "Information Society Technologies" Programme (1998-2002)

Abstract

The database of collected data by COSIN is divided in several sections. For the Internet section we collected traceroute data (2001-2002) and ping data (2001-2004), the data available can be considered a snapshot of a part of the Internet. For the World-Wide Web section we searched and analyzed more than 300000 web pages, looking for the URLs they contained and detecting communities. We have included in the Protein Network section the first protein network data taken from the Database of Interacting Proteins. In the database are also available a series of miscellaneous data such related to Food Webs, Social Networks and U.S. Patents.

Our future action during the third year will be to have, alongside the "data" database, a "link" database: an annotated repository of links to sites that provide high quality data about various kinds of networks.

1. The Database

The database has grown around the preliminary Internet (traceroute) data we have collected in 2001-2002. As mentioned below, they represent an almost single-view of the Internet. It has been recently shown formally, by some of us [Petermann and De Los Rios 2004], that single-views of the Internet are not reliable, which was already known, empirically, in the computer science community. The necessity of going to massively multi-view probes of the Internet goes beyond the COSIN capabilities. The group in Paris-Sud has developed, as a consequence, strong ties with large consortia that are providing such a multi-view effort. Some of their data have been made available in our database. Yet, this first case shows the shift of attitude that we are necessarily undertaking: rather than having a single repository of data, we are going to link more and more to sites where huge amounts of data are kept, usually by the same groups that are collecting them. This has clearly the advantage that the data are constantly updated and managed.

This is particularly true for other data in our database: for example biological data from protein interaction networks are becoming easily available to the public, and a number of organisations are working toward updated databases where each node and edge of the networks is commented and given a degree of reliability. These kinds of efforts are necessary to properly work with network data and, in general, need an expertise of the details of the domain that are not available within COSIN, where the effort is directed more toward the understanding of the large-scale properties of networks by means of statistical and algorithmic techniques.

Our future action during the third year will be to have, alongside the "data" database, a "link" database: an annotated repository of links to sites that provide high quality data about various kinds of networks.

2. Internet Data

The database contains the traceroute data collected by COSIN from October 2001 to July 2002 and ping data from December 2001 to March 2004. The traceroute data represent various snapshots on the Internet network taken from two different nodes with different degrees of depth (the allowed number of hops before interrupting the probe). Their analysis has been the focus of the diploma thesis of Simone Triglia. Despite this initial success, collection of large datasets about the structure of the Internet using traceroute has not been as fruitful as predicted. Indeed, since early 2002 it has been pretty clear that other consortia around the world were devoting many more resources to this task than the ones that COSIN could gather. In particular it has emerged that single-view data (data collected starting from a single node) can be highly skewed (this has been seen practically and analyzed theoretically, see [Petermann and De Los Rios 2004, Barrat et al. 2004]).

Our data, taken from two different sources, are essentially equivalent to a single-view probe). Therefore only massively multi-view data, beyond the capabilities of COSIN, can provide a faithful representation of the Internet. Such task is currently undertaken by the National Laboratory for Applied Network Research (NLANR, http://moat.nlanr.net) and by the Cooperative Association for Internet Data Analysis (CAIDA, http://www.caida.org), among others. COSIN, mainly through CR8, has established strong ties with some of these consortia. As a consequence, rather than collecting data ourselves, we have largely used their data for network analysis. Some of these data are available through the database, others are directly available from the above mentioned consortia.

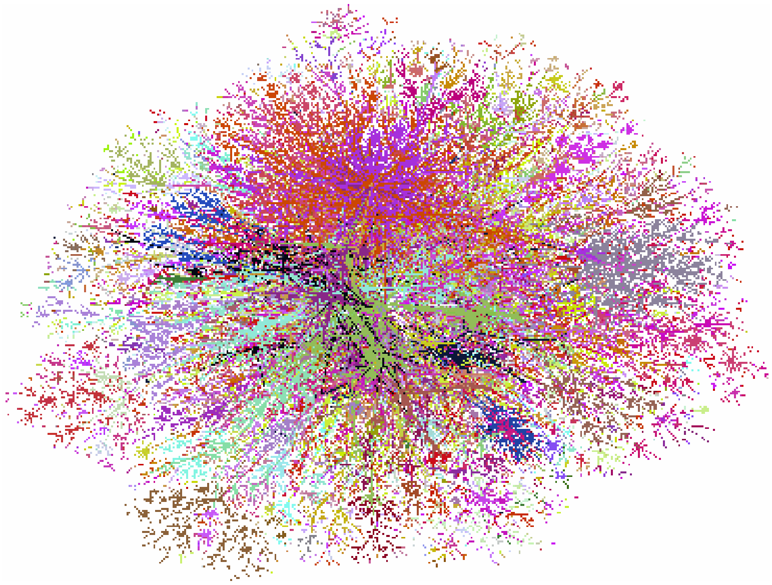

Figure 1 A typical map of the Internet as seen in the Internet Mapping Project.

The ping measures were taken from a machine in Milan that launched pings to several machines located around the world every half an hour. The aim was to gather a set of data to estimate the capacity of paths avoiding to use bulk data transfers that are considered an intrusive way measuring throughput and related quantities. This non-intrusive (or at least less intrusive) ways of guessing the capacity, based on the analysis of traceroute delays for varying packet size, or the so-called packet dispersion technique. The basic idea is as follows. Suppose packets are sent by host A to host B on a path consisting of links with capacities C1, C2, C3, …, Cn and assume for simplicity that there is no traffic on the path. The time it takes host A to transmit a packet of size d(bits) on the first link is d/C1 (sec). Thus, when two packets are sent back to back, the first bits of the two packets are separated by a time d/C1. Assume the second link is narrower than the first link (as is generally the case, since the first link is usually a LAN link). On the second link, the time separation between the first bits of the two packets will be d/C2 > d/C1. If the next link has capacity C3 > C2, the time separation of the first bits remains the same, but the packets will not be back to back anymore (there will be a gap between them equal to d(C3-C2)/C2C3). At the time of arrival at the host B, the time separation between the first bits of the two packets will be equal to d/Cmin, where Cmin is the capacity of the narrowest link on the path, i.e. the capacity of the path. Thus, one can in principle measure the capacity of a path by measuring first-bit time delays between packets. In the presence of traffic on the path, this nice deterministic argument breaks down and to make the technique work one has to resort to statistical arguments. There exist several concrete implementation of this idea, such as the software pathrate (by Dovrolis).

3. World-Wide Web Data

Two sets of web pages have been collected and analyzed during the end of 2003 and the beginning of 2004. First we used a software we developed (available in the software section) to query Google and get one thousand web pages containing a specific word. Then we used a second software we developed to :

- check if the page contained the specific word we chosen, and proceed only if it did
- count the outdegree of this page (the outdegree is the number of the links contained in a page)
- follow all the links of this page
- for each destination page check if it also contained the specific word we chose, and if so count the outdegree

Therefore each set of data available to the public is made by:

- a list of web urls we got from google, that actually contained a specific word. The outdegree of these pages.
- for each of these urls, the list of the web pages they pointed to, and containing the same word. The outdegree of these pages

More than 300000 web pages have been downloaded and checked to acquire these two data sets. They represent the web pages containing a word that are linked from web pages that contain the same word.

This set of data can turn out useful in studies related to the detection of communities in large networks. Some of the people involved in the COSIN project developed an algorithm to detect community structure in complex networks. The algorithm is based on spectral methods and takes into account weights and links orientations. Since the method detects efficiently clustered nodes in large networks even when these are not sharply partitioned, it turns to be especially suitable to the analysis of social and information networks.

4. Protein Networks Data

Protein networks are emerging as one of the new important fields of application of network theory, and as one of the most challenging test beds of current models. Actually, there is a circular feedback between protein and information networks as far as techniques of analysis are concerned.

There are thousands of different proteins active in a cell at any time. Many act as enzymes, catalysing the chemical reactions of metabolism. Others are components of cellular "machines", like ribosomes that reads genetic information and synthesises proteins.

Many other proteins are involved in the regulation of gene expression (protein production) in some way. There are proteins that play specific roles in special cellular compartments and others move from one compartment to another, acting as "signals". By directly interacting with one another, proteins continually affect one another's functions.

Proteins are produced and degraded all of the time. The rates at which these processes occur depend on what proteins are already present, how they interact with one another and how they interact with genes (DNA or messenger RNA). Proteins that bind to DNA or RNA often have a direct effect on the production or degradation of other proteins. One protein can speed up or slow down the rate of production of another by binding to the DNA or RNA (the genetic information) that is needed to make it.

The chart represents protein-protein interactions as lines (edges) forming a network between points (nodes). It doesn't show the functions of different proteins or the effect of the interactions. The relatively large number of red points in the chart demonstrates the fragility of cells. When a mutation causes the loss of one of these essential protein functions the cell dies. Correspondingly, the large number of green points in the chart demonstrates the robustness of organisms. The organism can survive loss of these proteins. Much of this robustness is due to "degeneracy" in the network - more than one protein, encoded in separate genes, serving the same or a closely similar function.

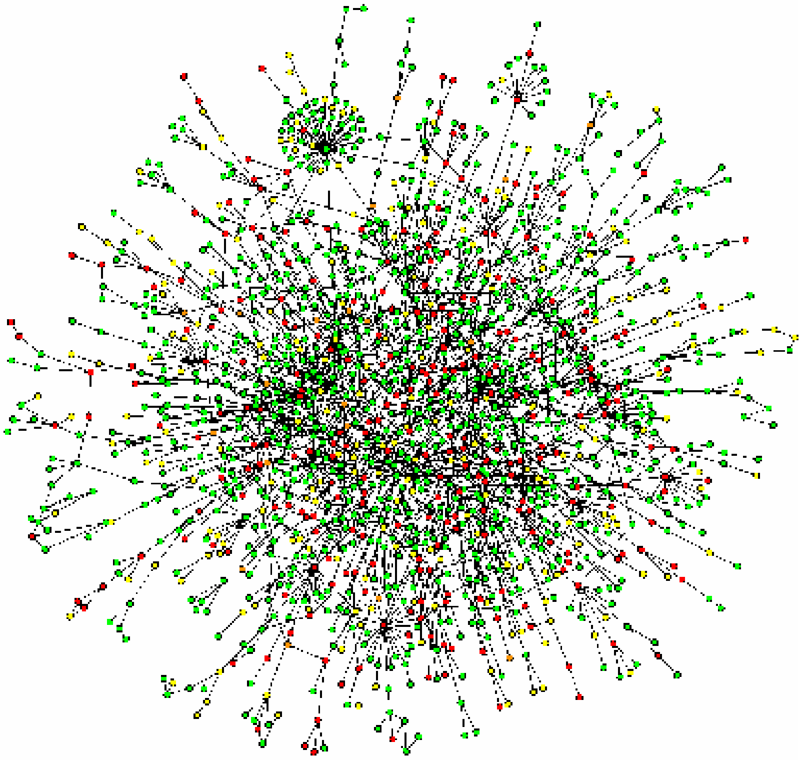

Figure 2 The Protein Interaction Map pf the Saccharomyces Cerevisiae (Baker's Yeast).

A complete map of protein interactions could certainly help in the understanding of the complex mechanisms running in a cell. Two proteins are said to interact if they bind together more than a certain elapse of time. Although long range interactions of the Van der Waals type are present, they are not accounted for.

The most popular and efficient method for recognizing if two proteins interact together is the *double hybrid* method. Let's say we would like to find out if protein A binds to protein B. We construct then an hybrid protein A and the analogous hybrid protein B, by attaching to protein A and B the two constituent parts of the *activator* of a certain protein C. If protein A anb B bind together, then the activator of C is reassembled and the production of protein C takes place. Usually, protein C is chosen among those proteins that are easily detectable. At the end, the presence of protein C indicates that proteins A and B interact each other. Unfortunately, with this double hybrid method, we collect no information about the strength of the interactions, so that the construction of weighted graphs is for the moment excluded.

We have included in the database the first protein network data taken from the Database of Interacting Proteins (DIP Database, http://dip.doe-mbi.ucla.edu/). This database is rapidly growing and contains the most extensive protein networks to date.

5. Miscellaneous Networks Data

Along with networks from well-defined contexts (Internet, WWW, Proteins) we have collected data about various systems. In particular the available data sets are:

- **Food Webs:** these data represent the structure of the ecological predator-prey relations in six ecosystems, collected in the years 1991-2001. Their analysis has been the object of one COSIN publication, where analogies between the optimization of transportation networks (and the Internet can be considered as one) and the principle of minimal resources dissipation have been drawn.
- **Social Networks:** the most celebrated social network is the actor collaboration network, that has been one of the first to be studied by the community. It can be obtained, in a format not easily readable, from the Internet Movie Database (http://www.imdb.com). Our database contains an elaborated data set that has been re-formatted so to be easily analyzed.
- **U.S. Patents:** there ample data about the network of patents in the U.S., available from the site http://www.nber.org/patents. We have provided a link to these data.

Papers analyzing the content of data repository

[Barrat et al. 2004] A. Barrat, A. Vazquez and A. Vespignani, Evaluating the statistical properties of unknown networks by merging partial spanning trees: a numerical study, in preparation 2004.

[Petermann and De Los Rios 2004] T. Petermann and P. De Los Rios, Exploration of Scale-Free Networks, Eur. Phys. J. B,.

[Capocci et al. 2004] Andrea Capocci, Vito D.P. Servedio, Guido Caldarelli, Francesca Colatori, Detecting communities in large networks (arXiv:cond-mat/0402499 v2 20 Feb 2004)

[Coccetti and Percacci 2003] Bandwidth measurements and router queues