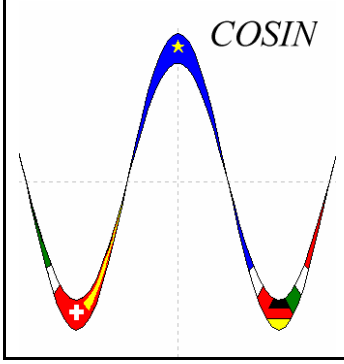

COSIN

IST–2001-33555

Coevolution and Self-Organization in Dynamical Networks

http://www.cosin.org

Statistical Analysis of Collected Data

Deliverable Number:	D20
Delivery Date:	September 2005
Classification:	Public
Partner Owning:	CR4 (UNIL-EPFL)
Contact Authors:	Paolo De Los Rios *Paolo.DeLosRios@epfl.ch*
Project Co-ordinator:	Guido Caldarelli (INFM) *Guido.Caldarelli@roma1.infn.it*

Partners:

CO1 (INFM) INFM *Italy*

CR2 (UDRLS) Università "La Sapienza" *Italy*,

CR3 (UB) Universitat de Barcelona *Spain*

CR5 (ENS) École Normale Supérieure, Paris *France*

CR7 (UNIKARL) Universität Karlsruhe *Germany*

CR8 (UPSUD) Université de Paris Sud *France*

CR9 (EPFL) Ecole Polytechnique Fédérale de Lausanne *Switzerland*

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

Abstract

The analysis of real data is an unavoidable part of the scientific process, since only real data can falsify models and direct the theoretical approach. Within COSIN we have analysed datasets from different domains, from technological to biological with the twofold goal of providing a benchmark for the modelling community (and for Deliverable D24), and to find hints of the *first principles* that govern the formation and evolution of large complex networks. Our results, starting from the preliminary analysis made in deliverable D5 from a general perspective, show that current models fall short of providing satisfactory descriptions of real systems and that topology based models will have, in the near future, to be modified in order to take into account the dynamics of the systems that are described using the network framework.

CONTENTS

INTRODUCTION

Every science advances through the close interaction of experiments and theory. In the case of large complex networks, it is not yet possible to perform extensive experiments at the scale of the whole systems. Therefore data must come from suitable probes able to observe the features of the networks at a global scale. Once large scale data have been obtained, their statistical characterization is the first step toward understanding the mechanisms that shape the formation of large networks, which in turn is an essential ingredient to formulate accurate models. There is therefore a natural overlap between the results of this deliverable and the ones of D24.

In this report the focus is on the findings rather than on the techniques and technicalities that have been used, which are nonetheless not less important, since the methods used to obtain large scale data could in principle deform the appearance of the real network, introducing systematic errors in the measurement process.

INTERNET DATA

We have used data from different sources, such as the National Laboratory of Applied Network Research (NLANR), of the Cooperative Association for Internet Data Analysis (CAIDA) and of the Topology Project at the University of Michigan.

Various authors have established the power-law distribution of the Internet degrees at different levels (IP, routers, Autonomous Systems) [1,2], and many models are known that are able to reproduce such feature. Yet Vázquez *et al.* [3] have shown that most of the same models are unable to reproduce higher level correlations found in real data. These results clearly points to the importance of a deeper characterization of the Internet through a thorough analysis of the data currently at hand.

By analyzing the Autonomous Systems network at different time frames (1997 to 1999), Vázquez *et al.* [3] found that its statistical properties are stationary, which is a first important result, since models are more easily solved in the stationary state. The typical power-law distributions of the degrees were re-obtained, along with the power-law distributions of the node betwenness. These results are easily compatible with most models. Yet, current models are mostly featureless as far as nearest neighbour correlations are concerned: the clustering coefficient c_k and the average connectivity of nearest neighbours, k_{nn}, of a node of degree k do not depend on k, whereas the same quantities on the AS network behave as $c_k \sim k^{-0.75}$ and $k_{nn} \sim k^{-0.5}$. Among the models checked in [3], only the fitness model of Bianconi and Barabási [4] showed

correlations that are *qualitatively* in agreement with the real data. We refer to the Report for "D24: Modelling the Internet graph" for a more extensive review of the models we have subsequently proposed to reproduce also nearest neighbour correlations.

The clustering coefficient of a node of degree k counts the number of triangles it belongs to, normalized by $k(k-1)/2$, which is the maximum number of triangles it could belong to. The clustering coefficient is a measure of the number of cycles of length three present in the network. The concept is therefore easily extended to longer loops. Caldarelli *et al.* [5] have shown that the number of quadrilaterals in real networks largely exceed, by one to four order of magnitudes, the results for random graphs with the same size, average degree and degree distribution. This is a clear signature of relevant correlations in real networks that must be recovered by any model. Bianconi *et al.* [6] have found that although the number of long length cycles in the Autonomous Systems network is larger than for random networks, it can be predicted by using simple nearest neighbour correlations: the AS network is therefore "markovian". This result is important both for a better social and economic understanding of the rules governing the AS network and for the modelling efforts: indeed it suggests that the unknown precise underlying mechanisms forging the AS network shape are likely to involve only considerations about pairs of nodes that become connected, disregarding the properties of second and farther away neighbours. Very recently Ángeles Serrano *et al.* [7] have proposed a model where some more detailed microscopic decision making behaviour is introduced, which is able to more closely reproduce the correlations observed in [6].

Analyzing Round-Trip-Times (RTT) of packets travelling over the Internet, Percacci and Vespignani [8] and Carbone *et al.* [9] have found that, once normalized by the geographical distance d of the source-destination pairs, they are power-law distributed. The quantity RTT/d is the velocity of the packet, which in turn represents a measure of the quality of the Internet performances. The power-law distribution is a clear indication that performances are very heterogeneous over the Internet: improving the Internet therefore does not amount solely to improving the average velocity, which would benefit only the "average" user, but also to reduce the statistical fluctuations, a result that would positively affect *all* users.

THEORETICAL DIFFICULTIES

The statistical analysis of data is important, but only if their reliability is assessed. Actually, it has been pointed out that the exploration itself of the Internet, based mainly on the *traceroute* command, can be biased toward scale-free networks even if the underlying network is not scale free [10,11], and that scale-free networks can see their *real* exponents changed [12]. The reason

for the skew in the data is that any algorithm exploring the network from a single source is likely to miss many edges, which affects in a more dramatic way low degree nodes that can be missed completely. As a consequence low degree nodes are underrepresented in the final degree distribution, to the advantage of its tail.

The only way to restore the real network topology is to measure the network from multiple sources. Rigorous analytical results [13,14] show that by suitably choosing the ratio of sources and destinations the bias in the results can be corrected. This result is of particular relevance for massive Internet measurement projects currently on-going. The DIMES project (part of FP6 Evergrow) is a large scale distributed exploration that should be able to implicitly implement the recipes found analytically [15].

ECONOMIC DATA

Modelling economic data has relied for years on a number of assumptions that a thorough analysis of the growing amount of available data has shown not to hold. Hence finding as many possible patterns as possible in real data is of paramount importance to falsify widespread financial models. A network representation of stocks and their holders represent one method, although of course not the only one, to frame information from real data.

Bonanno *et al.* [16,17] have analyzed the time-correlations between stocks in the New York Stock Exchange over a time window of 12 years, and then built the Minimal Spanning Tree (MST) of the stocks by connecting pairs of stocks according to a distance related to the correlations. The statistical properties of the MST from real data have been then compared to the ones from two commonly used models: uncorrelated Gaussian time series and the one-factor model. None of them is able to reproduce the same power-law distributions of the degrees and in-components of the nodes, clearly showing that a more refined modelling is needed that would start from premises that are in agreement with real data.

Battiston [18] has analyzed the shareholding networks of the New York Stock Exchange, of the Nasdaq and of the Italian MIB. A shareholding network is in principle a bipartite network where vertices are companies and shareholders, and an edge connects a shareholder to a company when she posses some stocks of that company. Actually, shareholders are most often other companies equally well present in the stock market, so that it is more natural to represent a shareholding network as a directed network with edges directed from holders to owned companies. The number of effective holders of a stock is defined as the number of holders that effectively control the company, and the effective number of stocks controlled by holders is the effective number of companies that they can influence through their shares of stocks. Defining such objective measures makes it possible to compare disparate stock exchanges, and discover their profound

differences: the US markets have a more spread shareholding network, made of one large connected component, whereas the Italian one is made of 89 separate components. Moreover most MIB listed companies are controlled by a single holder, against the average six holders of the US companies. These results imply some tantalizing economic, social and historical considerations that go of course beyond the scope of the paper's analysis. A further result coming from the analysis of the shareholding network is that the sizes of stock portfolios are power-law distributed and that the diversification of portfolios depends on their sizes as an algebraic function [19]: larger portfolios contain more different stocks. Again, this empirical finding is at odds with financial models that posit that portfolios can be indefinitely diversified irrespective of their size.

Caldarelli and Catanzaro [20] have analyzed the social network of corporate boards. It is a bipartite network of directors and companies, where an edge is present between a company and a director if he/she sits in its board. It is then possible to construct the social network of directors, where two directors are connected if they sit in at least one common board. Not unexpectedly, the resulting network is small-world and with *strong* links, implying that some directors sit together in more than a single board. The consequences of this finding for the spread of insider information cannot be underestimated.

Overall, these findings suggest on the one hand that many economic mathematical models currently widely used rest on wrong assumptions, and that a greater effort should be made toward a more realistic description of stock markets; on the other hand the structure of the shareholding and directors boards networks suggest the presence of different socio-economical mechanisms behind the US and Italian stock markets.

ECOLOGICAL DATA

Networks are finding growing application in the modelling of ecological data. The predator-prey relations that define food-webs are best expressed as edges of networks, and the taxonomic relations between species, genera, and so on are naturally represented as loop-less networks, that is, trees.

Garlaschelli *et al.* [21] have analyzed the topological properties of 7 well characterized food-webs, finding that they are optimized to efficiently transport energy (bio-mass) from the lowest to the highest trophic levels, approximately identified with herbivores and top predators respectively, although this kind of cartoon representation might be misleading. More importantly, through the definition and measurement of suitable topological quantities, they have been able to show that all the 7 food-webs behave in the same way, pointing to a universal

outcome for the optimization process. These results have sparked a lively debate mainly due to the scarcity of available data [22], which will for sure prompt ecologists to improve the quality of their field observations.

Caretta Cartozo *et al.* [23] have studied the statistical properties of the taxonomic trees of the *florae* of different bio-geographical regions. Regardless of geographic location, climate and environment, all collections have universal statistical properties, different from the ones of a random sample of the same number of species taken from a pool of species from all over the world. These result show again that the final outcome of the organization of ecosystems is likely universal, which in turn prompts the search for the underlying mechanisms behind the self-assembly of a working collection of species.

Caldarelli *et al.* [24] have also shown that one of the properties that had been hailed as universal, namely the inverse square-law distribution of the sizes *s* of taxonomic levels at all levels, is the trivial consequence of the tree-like organization of taxonomic trees. Whether the trees come from real ecosystems or they are random collections of species, the tree-like structure inherent to hierarchical taxonomy leads to the inverse square law. Minimal and Random Spanning Trees built on random networks and on networks grown according to different models (see Report for D24) also show an inverse square distribution of the sizes of the sub-trees at all levels. These last results highlight the importance of a good choice of the topological quantities to analyze, since some of them do carry relevant information, whereas others do not.

WEIGHTED NETWORKS DATA

The topology of networks has received most of the attention over the years mainly because of the lack of data for the values, or *weights*, of the nodes and edges. Recently Barrat *et al.* [25] have had the opportunity to look to the weights of two networks, namely the world-wide airport network (WAN) and the scientist collaboration network (SCN), whose topological scale-free nature had already been established. The weights are more easily defined on the edges of the network: in the WAN the weight of an edge is the number of passenger seats available to travel with direct flights between the two airports connected by that edge; in the SCN the weight of an edge connecting two scientists is the number of co-authored papers. The weight of a network node (its *strength*) is then simply given by the sum of the weights of the edges it belongs to.

The authors' findings are that also the weights and strengths of the network are power-law distributed, and that there are correlations between the nodes' degrees and their strengths. In the case of the SCN strengths and degrees are simply proportional, whereas in the WAN there is a non-trivial power-law relation.

All these results have called for new models that couple the evolution of the topology to the weights' dynamics [26,27] (see Report for D24).

CONCLUSIONS

Our efforts for this deliverable, in connection with D24, have been as much as possible to link the statistical analysis of data to the formulation of theoretical models. We have been able to show that, in many cases, current models are not satisfactory (see Report for D24) because either they do not reproduce the observed quantities or because even the model premises do not hold, as is the case of many financial models.

At the same time it is important to stress that data must not be taken directly by face value: some COSIN members, and some other authors, have been able to show that the exploration methods, that is, the way vertices and edges are detected, can skew the data so that the observed topology can be close to a scale-free one even if the real network is not.

The results for weighted networks, moreover, clearly point to a co-evolution of the topology and of the intrinsic features of nodes and edges, and future models will have to take this further ingredient into consideration.

BIBLIOGRAPHY

[1] M. Faloutsos, P. Faloutsos and C. Faloutsos, *On power-law relationship of the Internet topology*, Comput. Commun. Rev. **29**, 251-263 (1999).

[2] G. Caldarelli, R. Marchetti and L. Pietronero, *The fractal properties of the Internet*, Europhys. Lett. **52**, 386-391 (2000).

[3]• A. Vázquez, R. Pastor-Satorras and A. Vespignani, *Large-scale topological and dynamical properties of the Internet*, Phys. Rev. E **65**, 066130 (2002).

[4] G. Bianconi and A.-L. Barabási, *Competition and multiscaling in evolving networks*, Europhys. Lett. **54**, 436-442 (2001).

[5]• G. Caldarelli, R. Pastor-Satorras and A. Vespignani, *Structure of cycles and local ordering in complex networks*, Eur. Phys. J. B **38**, 183-186 (2004).

[6]• G. Bianconi, G. Caldarelli and A. Capocci, *Loops structure of the Internet at the Autonomous System level*, Phys. Rev. E **71**, 066116 (2005).

[7]• M. Ángeles Serrano, M. Boguñá and A. Diaz-Guilera, *Competition and adaptation in an Internet evolution model*, Phys. Rev. Lett. **94**, 038701 (2005).

[8]• R. Percacci and A. Vespignani, *Scale-free behavior of the Internet global performance*, Eur. Phys. J. B **32**, 411-414 (2003).

[9]• L. Carbone, F. Coccetti, P. Dini, R. Percacci and A. Vespignani, *The spectrum of Internet performance*, in Proceedings of Passive and Active Measurement (PAM2003).

[10] A. Lakhina, J.W. Byers, M. Crovella and P. Xie, *Sampling biases in IP topology measurements*, Technical report BUCS-TR-2002-021, Department of Computer Science, Boston University.

[11] A. Clauset and C. Moore, *Accuracy and scaling phenomena in Internet mapping*, Phys. Rev. Lett. **94**, 018701 (2005).

[12]• T. Petermann and P. De Los Rios, *Exploration of scale-free networks: Do we measure the real exponents?*, Eur. Phys. J B **38**, 201-204 (2004).

[13]• L. Dall'Asta, I. Alvarez-Hamelin, A. Barrat, A. Vázquez and A. Vespignani, *Statistical theory of Internet exploration*, Phys. Rev. E **71**, 036135 (2005).

[14]• L. Dall'Asta, I. Alvarez-Hamelin, A. Barrat, A. Vázquez and A. Vespignani, *Traceroute-like exploration of unknown networks: a statistical analysis*, Lecture Notes in Computer Science **3045** (2005) and cond-mat/0406404.

[15] www.netdimes.org

[16]• G. Bonanno, G. Caldarelli, F. Lillo and R.N. Mantegna, *Topology of correlation-based minimal spanning trees in real and model markets*, Phys. Rev. E **68**, 046130 (2003).

[17]• G. Bonanno, G. Caldarelli, F. Lillo, S. Micciché, N. Vandewalle and R.N. Mantenga, *Networks of equities in financial markets*, Eur. Phys. J. B **36**, 363-371 (2004).

[18]• S. Battiston, *Inner structure of capital control networks*, Physica A **338**, 107-112 (2004).

[19]• D. Garlaschelli, S. Battiston, M. Castri, V.D.P. Servedio and G. Caldarelli, *The scale-free topology of market investments*, Physica A **350**, 491-499 (2005).

[20]• G. Caldarelli and M. Catanzaro, *The corporate board networks*, Physica A **338**, 98-106 (2004).

[21]• D. Garlaschelli, G. Caldarelli and L. Pietronero, *Universal scaling relations in food webs*, Nature **423**, 165-168 (2003).

[22] Garlaschelli *et al. reply*, *Universal scaling in food-web structure*?, Nature **435**, E4 (2005).

[23]• C. Caretta Cartozo, D. Garlaschelli, C. Ricotta, M. Barthelemy and G. Caldarelli, *Quantifying the universal taxonomic diversità in real species assemblage*, submitted (2005).

[24]• G. Caldarelli, C. Caretta Cartozo, P. De Los Rios and V.D.P. Servedio, *Widespread occurrence of the inverse square distribution in social sciences and taxonomy*, Phys. Rev. E **69**, 035101 (2004).

[25]• A. Barrat, M. Barthelemy, R. Pastor-Satorras and A. Vespignani, *The architecture of complex weighted networks*, Proc. Natl. Acad. Sci. USA **101**, 3747-3752 (2004).

[26]• A. Barrat, M. Barthelemy and A. Vespignani, *Weighted evolving networks: coupling topology and weight dynamics*, Phys. Rev. Lett. **92**, 228701 (2004).

[27]• A. Barrat, M. Barthelemy and A. Vespignani, *Modelling the evolution of weighted networks*, Phys. Rev. E **70**, 066149 (2004).

Papers marked with the • symbol have been produced with the acknowledged support of COSIN.