COSIN

IST–2001-33555

Coevolution and Self-Organization in Dynamical Networks

http://www.cosin.org

Modelling the Internet Graph

Deliverable Number:	D24
Delivery Date:	15 April 2003
Classification:	Public
Partner Owning:	CR9
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

Finding a model that could reliably describe the Internet's structure and the principles shaping it would be a precious result, since it would open the possibility to play with the Internet, to see the effects of different kind of perturbations and, ultimately, to try designing a better network. Yet, this task is hindered by our ignorance of the microscopic mechanisms at work in shaping the Internet's structure. This is a particular application of the most general issue on the analysis of collected data (D19). Here we use the analysis of Internet data as a benchmark against which any model that is proposed has to be tested. This has been our approach in this deliverable, and we have found that a purely topological description is very likely insufficient to describe the Internet, and that some further, finer level of details, capturing the intrinsic qualities of nodes and edges, should be included in the models to go beyond simplistic self-referential topological mechanisms.

CONTENTS

INTRODUCTION

Modelling the Internet graph is a challenge that is inextricably intertwined with the statistical analysis of Internet data (deliverable D20) so that a certain degree of overlap between these deliverables is unavoidable.

The Internet graph can be analyzed at different levels (IP, routers, Autonomous Systems), each with its own peculiarities. As such, the ideal approach would be to have a single model able to reproduce all the measured topological network quantities at all levels. This is at present unrealistic because the evolution of each level may well be governed by different principles that are mostly unknown: as such the modelling of the Internet graph has to rely on simplistic models defined by a few, paradigmatic rules.

Network growth: The Internet (at all levels) changes over time by the continuous addition/deletion of nodes. Thus, models for the Internet should either directly describe the network dynamics or be compatible with it, meaning that the same outcome should be obtained irrespective of considering all nodes as present from the beginning or added/deleted one by one.

The classical example of an intrinsically dynamical network model is the Barabási-Albert model [1], where new nodes enter the network at a constant rate and choose older ones as possible connection partners. Clearly *older* nodes behave differently than *newer* ones, so that their topological properties bear a record of their history. The Erdös-Rényi model [2] is instead an example of a model where the final result does not change whether its nodes have been added over time or they have been considered as present from the start.

Connection criteria: when entering the Internet, new nodes, whose identity depends on the level at which the Internet is analyzed and modelled, choose their connections according to specific criteria that are a necessary ingredient of any Internet model. However, the rules of the model clearly cannot capture the precise decision mechanisms at work in the real Internet. Rather, they are cartoon representations of real processes, simple yet general enough that they capture the main ingredients of the real system.

DYNAMIC MODELS

The first observed property of real world complex networks that is not possible to describe using Erdös-Rényi (ER) networks is the distribution of the number of connections (the degree) of the network vertices, which is a Poisson function for ER networks whereas in the Internet and other real networks it is a power-law with diverging second moment (hence their name of scale-free, SF, networks). The model of Barabási and Albert (hereafter BA model) has shown that SF networks can emerge from two simple rules: new nodes are added to the network at a constant rate and they preferentially choose, as connection partners, nodes with already a large value of their degree (*preferential attachment*, PA). Although the BA model has been instrumental to go beyond ER networks, its outcomes do not satisfy many of the properties of real networks. Indeed, as recently shown by Vázquez *et al.*[3], Caldarelli *et al.*[4] and Bianconi *et al.*[5], the Internet is characterized by non-trivial correlations at the local level, which the BA model is unable to reproduce. Vázquez *et al.*[3] have further shown that also Molloy-Reed networks [6] (where networks are built by randomly connecting nodes whose degrees are taken from a probability distribution similar to the Internet's one) and the generalized BA model (GBA)[7] cannot capture the Internet's correlations. Only the fitness model of Bianconi and Barabási [8] shows correlations qualitatively similar to the observed ones.

In the BA model new nodes are added at a constant rate and connected with older ones according to the preferential attachment rule. In the GBA model existing links are furthermore rewired at a constant rate and still in accord with preferential attachment: a node is chosen and one of its edges is rewired to another target chosen with a probability proportional to its degree. A different rewiring mechanism, inspired by the GBA, has been proposed by Catanzaro *et al.*[9] and by Caldarelli *et al.*[10]: at every time step, with probability *p* a node *i* is chosen and one of its edges is rewired to a new target *j* chosen with a probability which is a function $f(|k_i-k_j|)$, where k_a is the degree of node a. Thus rewiring takes place only after that node *i* has compared the other node's degree with its own. In [9] the rewiring probability is a decreasing functions of $|k_i-k_j|$ so that nodes of similar degree are more likely to connect to each other. The outcome is a class of *assortative* networks, where high degree nodes are likely to be connected with each other. These networks do not reproduce the *disassortativity* (that is, high degree nodes are unlikely to be connected to each other) of the Internet's correlations at the Autonomous System level, but they show that when the degree difference ingredient is taken into account, correlations can naturally appear. Disassortativity can be recovered by taking $f(|k_i-k_j|)$ as a growing function of its argument [10]: the largest the difference between the degrees, the higher the rewiring probability. Although this result, in light of the preceding one, is not completely unexpected, the more interesting result is that as the intrinsic probability of rewiring $p\tilde{}0.5$, almost all the correlations measured for the Internet (at the AS level) are recovered to a good quantitative measure. This result is not a proof that the model proposed in [10] does capture the real mechanisms that forge the Internet. Rather it points out that nodes likely establish and mutate connections in time only after having compared some properties of the possible partners with their own.

Vázquez *et al.*[3] showed that the Bianconi and Barabási fitness model (BB) [8] is able to qualitatively reproduce the correlation trends in the Internet. The BB model proposes a modification of the basic PA rule by assigning an intrinsic value (the *fitness*) to the network nodes, in the form of a variable taken from a given probability distribution. A new node entering the network chooses its partners with a probability that is proportional to the product of their degree and of their fitness. Yet, no comparison between the fitness of the new node and of the older ones determines the choice. Capocci *et al.*[11] proposed a mechanism where the above mentioned fitness modified PA rule takes place only over nodes whose fitness is larger than the one of the new entering node, mimicking a choice where only intrinsically more authoritative nodes are considered as viable partners. The main result of this paper is that this model is able to reproduce, with a good approximation, the values of the correlations measured for the Internet.

A better quantitative agreement is recovered by a new model introduced by Ángeles Serrano *et al.*[12]. This model introduces a better *microscopic* description of the nodes of the Internet, by taking into account the hierarchical organization in users and Autonomous Systems. Moreover, also the dynamics of the vertices and edges is richer: both new users and AS join the network, although at different rates, users are allowed to change providers and AS can adapt the number of their edges so to satisfy the connectivity demands of their users. Each part of the dynamics being characterized by specific rates, partly fitted to the empirical growth rates of the real Internet, the network of Autonomous Systems turns out to be scale-free and the power-law degree distribution decays with an exponent close to 2, as measured in real data. Starting from these premises, Ángeles Serrano and co-workers then refine their model placing AS in two-dimensional space in such that they cover a fractal set, as measured for the Internet [13]. Connections between AS take then place according to the same rules as above, but also taking into account the costs of long-distance connections and the bandwidth needs of different AS (that is, only AS that need to increase their bandwidth because of an increase of users can link to each other). The final result is a model where very many microscopic details close to the ones of the Internet have been introduced and that is able to closely reproduce many of its correlation properties. Far from being an arrival point, the model of Ángeles Serrano and co-workers shoes that more and more details have to be plugged back in the models to obtain detailed a reproduction of the real data. At the same time it builds on and gives full credit to older, well assessed principles such as growth and preferential attachment, showing how to incrementally step up the amount of details.

The results presented in this section suggest that more than a single network generating mechanism can generate networks with local correlations in quantitative agreement with the observed ones. Although apparently non-satisfactory, this conclusion also calls for a better quantitative characterization of the Internet, and for a deeper understanding of the detailed, engineering, economical and possibly political mechanisms shaping the Internet. Without such interdisciplinary understanding we are bound to propose models and unfortunately more than a single one is able to reproduce the data available to date.

STATIC MODELS

Tangmunarunkit *et al.* [14] proposed that the topology of the Internet, at least at the AS level, could be determined more by intrinsic factors (the size of AS, in turn related to the technology and capital of the companies and agencies managing the various AS) than by a growing mechanism. For example, AS could link to each other only when their economical and technological gain is estimated to be large enough. Also geographical constraints, such as the distance between nodes, could affect the choices of connecting partners. These considerations are in agreement with the notion outlined above that nodes choose their connection partners only after a due comparison of the other nodes' properties and their own.

In order to single out the possible outcomes of models driven by the nodes' intrinsic properties, Caldarelli *et al.* [15] have completely neglected the growth ingredient, focusing rather on static networks. A number x (the fitness) drawn from a probability distribution $r(x)$ is assigned to every node of the network, mimicking in a very simplified way some intrinsic features. Nodes i and j are then connected to each other according to a probability p_{ij} that can depend on the fitness of both nodes, $p_{ij}=f(x_i,x_j)$. Interestingly this model defines a class of models that trivially includes the ER model as a special case, $p_{ij}=p$. Various choices of $r(x)$ and of $f(x_i,x_j)$ have been explored in [15]. A trivial choice for $r(x)$ is an algebraically decreasing function of x, $r(x)\sim x^{-a}$, and $p_{ij}=f(x_i,x_j)\mu\ x_i x_j$, which generate networks with a power-law degree distribution characterized by the same decaying exponent a. If x represents for example the size of the AS company then an algebraically decreasing distribution of x is in agreement with the empirical Zipf law of the sizes of companies, giving such a simple model a seemingly realistic interpretation. In [15] it has also been shown that power-law degrees distributions can emerge also from less intuitive fitness distributions and connections probabilities. In a series of papers it has been further proved that it is possible to use these simple network generating mechanisms to create graphs with desired degree *and* local correlation properties [16-18]. Finally a model where fitness are defined as arrays of random numbers has been introduced, opening the possibility to capture the intrinsic properties at a finer level of detail [19].

As previously stated, these models are static, whereas the Internet is intrinsically dynamic. Nonetheless the networks generated from these models, just as for the ER, can be allowed to grow by the constant rate addition of new nodes, each one characterized by its fitness. Since the connection probability does not depend on the time at which nodes have entered the network, considering the nodes as all present at the beginning or as added one by one does not change the final result.

WEIGHTED NETWORK MODELS

Although most of the above models do depend on some non-topological variables (the fitness) that can be considered as *weights* over the network, the term *weighted* network has been awarded only recently to a specific class of models where the non-topological quantities co-evolve with the topological ones.

It has always been widely accepted since the beginning of complex network research, in the late '90s, that networks should be characterized in a detail finer than their bare topology, and edges and vertices should carry tags describing their intrinsic, and possibly dynamically changing, properties (their weights). Yet, only recently some reliable and large scale weight data, specifically about scientific collaborations and airport traffic, have become available [20], which explains why the modelling of weighted networks has lagged behind for so long.

Vespignani and coworkers have recently proposed, in a series of papers [21-23], that the evolution of complex networks should also take into account the weights of vertices and edges in the connection choice process, and the most straightforward option is to substitute the topology based preferential attachment rule with a weight based one: in this way the weights determine the topology evolution. As new nodes enter the network they also affect the pre-existing weights, which need therefore to be suitably updated. The resulting model defines a dynamics where the topology is "slaved" to the weights.

This new class of models can be solved analytically to find that *both* the weights and most of the topological features of the corresponding networks are distributed according to power-laws, just as seen in real data. Still just as in every first modelling attempt, not all correlations turn out right. In the Worldwide Airport Network (WAN), for example, the weight of airports, measured as the number of passenger seats leaving from them, is an algebraically growing function of their topological degree. The basic models for weighted networks [21] are unable to explain such properties, and a number of alternatives have already been proposed [24], although none of them can yet fully and convincingly capture the precise observations.

Before exporting the above mentioned models to the Internet, a suitable definition of the weights is needed. For scientific collaboration networks weights are the numbers of co-authored papers; as mentioned, in the WAN weights are identified as the number of passenger seats along certain routes. Internet weights could be naturally chosen as the volumes of data packets travelling along the lines. Unfortunately large data sets of weights are still missing for the Internet, also because of strong industrial interests in keeping them private. Still, the above models, that already have had a strong impact on the community, can be applied, with possible suitable adaptations, once such data will be available.

CONCLUSIONS

Recent years have seen many different models able to reproduce qualitatively, and sometimes quantitatively to a good approximation, the basic topological features of the Internet, including some local correlations. Yet, such models are sometimes completely different, so that a seeming ambiguity is present. Actually modelling the Internet faces a daunting task: modelling a system in a reliable and definitive way needs a thorough understanding of the basic, *microscopic*, mechanisms and *degrees of freedom* of the system, that can then be cast in a suitable mathematical form able to transform such information into a quantitative predictive tool. Unfortunately the Internet is far from being well characterized at a microscopic level: the real market, technical and political criteria driving connection choices are largely unknown, and likely extremely heterogeneous over the network. As such it is at the moment unfeasible to have a quantitative predictive and reliable model.

Rather, we are bound to focus to what is measurable and use it as a *negative* tool, to decide which models clearly do not reproduce the observed behaviour of the Internet. Moreover, we cannot really single out one of the models that are up to the task against the others.

Our work for this deliverable as therefore been to try and understand what mechanisms go in the right direction, at least as far as a comparison with the available data is concerned. In doing so we have uncovered, though very different models, that a central ingredient that ought to be included in any future Internet model is the comparison that a node performs of its intrinsic properties with the ones of the potential connection partners. This ingredient has been shown to lead very easily to the same correlation trends observed in the real data.

Another problem hindering the direct applicability of more refined models, such as weighted network models, is the unavailability of large sets of traffic data, often because they are proprietary. Possibly the advent of Internet2, an academically managed high speed network, will make such data available to researchers allowing some more finesse in the modelling.

BIBLIOGRAPHY

[1] A.-L. Barabási and R. Albert, *Emergence of scaling in random networks*, Science **286**, 509-511 (1999).

[2] P. Erdös and P. Rényi, *On random graphs*, Publicationes Mathematicae **6**, 290-297 (1959).

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

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

[6] M. Molloy and B. Reed, *A critical point for random graphs with a given degree sequence*, Random Struct. Algorithms **6**, 161-179 (1995).

[7] R. Albert and A.-L. Barabási, *Topology of evolving networks: local events and universality*, Phys. Rev. Lett. **85**, 5234-5237 (2000).

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

[9]• M. Catanzaro, G. Cadarelli and L. Pietronero, *Assortative model for social networks*, Phys. Rev. E **70**, 037101 (2004).

[10]• G. Caldarelli, P. De Los Rios and L. Pietronero, *Generalized network growth: from microscopic strategies to the real Internet properties*, cond-mat/0307610.

[11]• A. Capocci, G. Caldarelli and P. De Los Rios, *Quantitative description and modeling of real networks*, Phys. Rev. E **68**, 047101 (2003).

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

[13] S.H. Yook, H.W. Jeong and A.-L. Barabási, *Modeling the Internet's large-scale topology*, Proc. Natl. Acad. Sci. USA **99**, 13382-13386 (2002).

[14] H. Tangmunarunkit, J. Doyle, R. Govindan, S. Jamin, S. Shenker and W. Willinger, *Does AS size determine degree in AS topology?*, Comput. Commun. Rev. **31**, 7-10 (2001).

[15]• G. Caldarelli, A. Capocci, P. De Los Rios and M.A. Muñoz, *Scale-free networks from varying vertex intrinsic fitness*, Phys. Rev. Lett. **89**, 258702 (2002).

[16]• V.D.P. Servedio, G. Caldarelli and P. Buttà, *Vertex intrinsic fitness: how to produce arbitrary scale-free networks*, Phys. Rev. E **70**, 056126 (2004).

[17]• M. Boguñá and R. Pastor-Satorras, *Class of correlated random networks with hidden variables*, Phys. Rev. E **68**, 036112 (2003).

[18]• M. Catanzaro, M. Boguñá and R. Pastor-Satorras, *Generation of uncorrelated random scale-free networks*, Phys. Rev. E **71**, 027103 (2005).

[19]• M. Boguñá, R. Pastor-Satorras, A. Díaz-Guilera and A. Arenas, *Models of social networks based on social distance attachment*, Phys. Rev. E **70**, 056122 (2004).

[20]• 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).

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

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

[23]• M. Barthelemy, A. Barrat, R. Pastor-Satorras and A. Vespignani, *Characterization and modeling of weighted networks*, Physica A **346**, 34-43 (2004).

[24]• A. Barrat, M. Barthelemy and A. Vespignani, *The effects of spatial constraints on the evolution of weighted complex networks*, J. Stat. Mech. P05003 (2005).

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