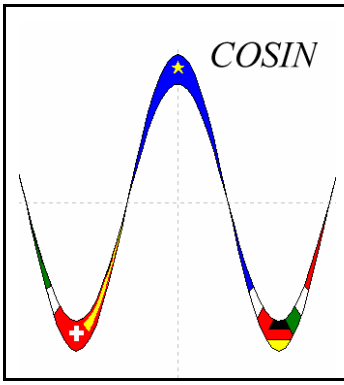


COSIN

IST–2001-33555

Coevolution and Self-Organization in Dynamical Networks

http://www.cosin.org

Papers and Book on Complexity

Deliverable Number:	D26
Delivery Date:	September 2005
Classification:	Public
Partner Owning:	C01 (INFM)
Contact Authors:	Guido Caldarelli *Guido.Caldarelli@roma1.infn.it*
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

Despite that COSIN project is focussed on the technological networks of Internet and the WWW, scale-free networks are present in several other fields of interest as those of biology, social science and genomics. All the applications in this field concur to the formation of the new science of complexity and ultimately help the progress in the understanding of technological networks. This deliverable is devoted to illustrate these studies on complexity and the summary that has been made in a special book on COSIN activity that is in press for the World Scientific Editor. The cover of this book together with the contents of the draft manuscript are reported in the Appendix.

COMPLEX NETWORKS

According to one commonly accepted definition, a system is *complex* when it cannot be split into smaller components whose behaviour and function can be understood individually, to be then re-assembled at a later stage. Rather, a complex system should be studied as a whole. Networks are therefore well suited to representing complex systems, since their interconnectivity already visually signals the difficulty of breaking them apart.

Since the connectivity patterns are paramount to understanding a network's behaviour, we should first ensure that the graph representation of the real system is a faithful one. A thorough analysis of the nodes and edges detection methods, and of their possible pitfalls, is thus necessary.

In a series of papers, we have shown that some data mining techniques can indeed skew the measurements introducing large systematic errors.

If the measurement is performed using the edges of the networks to carry some probe signals, such as the *traceroute* command in the Internet, the large scale statistical distributions of the network topological quantities can appear rather different from the real ones: random Poisson graphs can turn into scale-free graphs [1,2], and scale-free graphs can change their characteristic exponents[3], because low degree nodes are more difficult to detect, whereas the nodes in the tail of the distribution will be almost faithfully represented due to their high degrees. A theoretical analysis of this issue has led Dall'Asta *et al.*[4,5] to find the optimal ratio of source/destination pairs so to minimize the number of probes still capturing the real large scale structure of the network. These results advance in parallel with new distributed measurement efforts from various consortia such as DIMES (www.netdimes.org) and PlanetWeb (a consortium of more than 200 universities worldwide that should become public in the next few months) that indeed aim at getting a multi-source view of the Internet.

Caldarelli *et al.*[6] and Petermann and De Los Rios [3] showed that a protein interaction graph can appear as scale-free if it is only possible to discover edges representing interactions between proteins that are stickier than a certain threshold, which is a stated property of present detection methods, and also practitioners of the field have realised that the measurement method itself can heavily distort the resulting graph [7].

These results stress that the reliability of the data, although clearly a fundamental issue any field of science, has been a focus of research only very recently in the field of complex systems, and COSIN members have been playing a central role in these developments.

[1] 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.

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

[3]• 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).

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

[5]• 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.

[6]• 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).

[7] J.-D.J. Han, D. Dupuy, N. Bertin, M.E. Cusik and M. Vidal, *Effect of sampling on topology predictions of protein-protein interaction networks*, Nature Biotechnology **23**, 839-844 (2005).

EPIDEMIC SPREADING

Traditionally, mathematical models for the spread of a disease have relied on differential equations, describing the dynamics of spreading within uniformly mixed populations. The basic premise of uniform mixing is that all individuals in a group are equally likely to become infected. The spreading process itself is then captured using compartments, i.e. individuals belonging to epidemiological classes such as susceptible (S), exposed (E), infective (I) and recovered (R), between which the flows of population are described with rate equations. Within this framework, the simplest model is the widely-utilized SIR (Susceptible-Infected-Removed) model in which susceptible individuals may become infected and continue to infect others until finally removed from the system due to recovery, death, or containment. When an epidemics spreads over a network (of individuals, computers *etc.*), the large scale topology can play a major role in the velocity of diffusion of the pathogen agent and in its prevalence. A fundamental result has been obtained by Pastor-Satorras and Vespignani [8], who showed that the global scale-free property of graphs makes them more sensitive to epidemic spreading and confer to the pathogen a larger resilience.

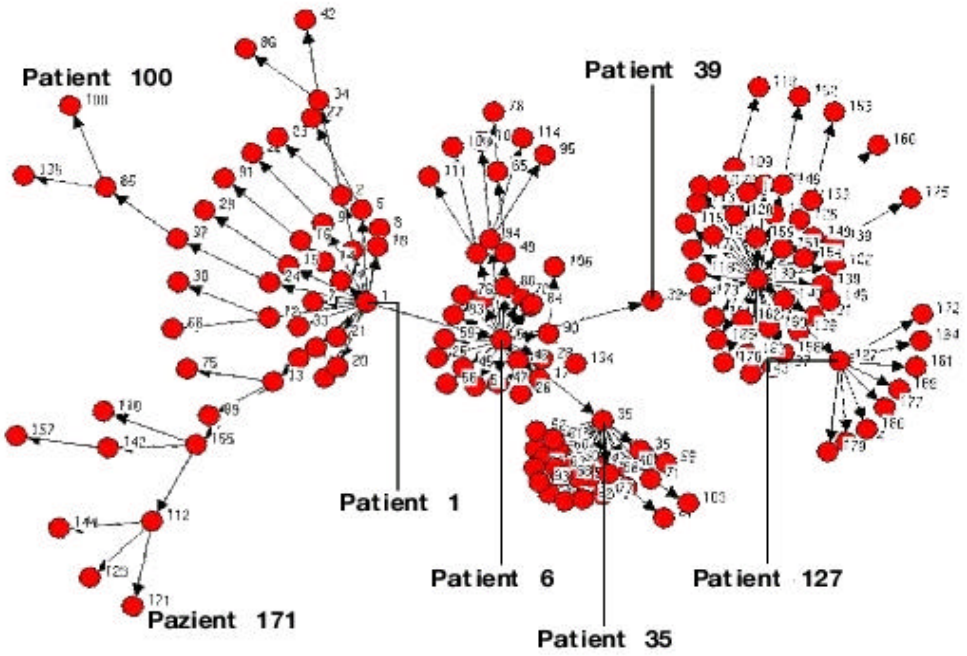


Figure 1: The scale free network of SARS infection

The local properties have been instead investigated by Petermann and De Los Rios [9,10], who, by developing new numerical approaches to epidemic spreading, have shown that different networks with the same average degree but different levels of local connectivity behave differently, with more locally interwoven networks being more resistant to infection.
These results show that in a complex system both global properties (being scale free) and local ones (local connectivity) significantly affect the system behaviour. This is further evidence that a complex system has to be studied taking into account many different scales at once.

[8] R. Pastor-Satorras and A. Vespignani, *Epidemic spreading in scale free networks*, Phys. Rev. Lett. **86**, 3200-3203 (2001).

[9] T. Petermann and P. De Los Rios, *Role of clustering and gridlike ordering in epidemic spreading*, Phys. Rev. E **69**, 066116 (2004).

[10] T. Petermann and P. De Los Rios, *Cluster approximations for epidemic processes: a systematic description of correlations beyond the pair level*, J. Theor. Biol. **229**, 1-11 (2004).

- **Food Webs**

Another typical case of networks in nature is represented by the Food Webs. In food webs we have that the various elements of an ecosystem represent the vertices of a graph whose edges are the predation relationships. Apart the ubiquitous scale-free distribution for the degree [11] indicating that evolution selected some organisms able to predate on very different species, one of the most interesting question is related to investigate if this kind of network has been shaped by evolution in order to optimise some cost function.

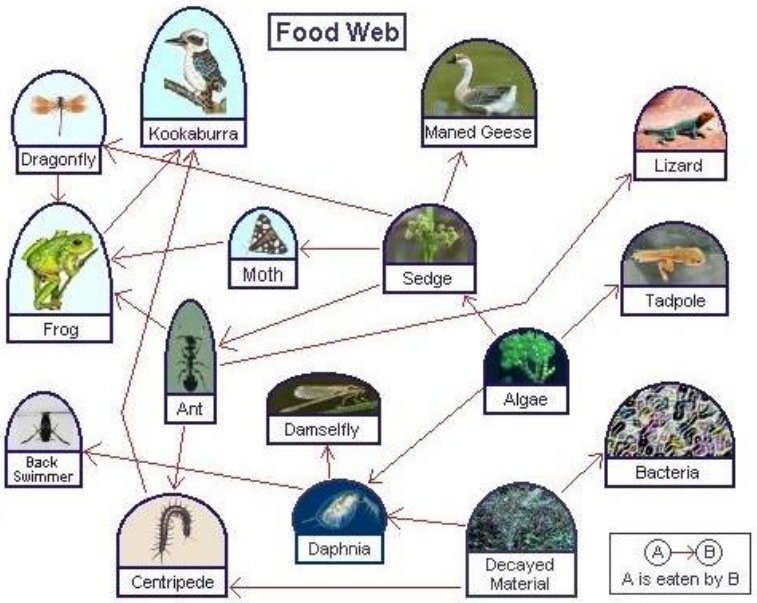


Figure 2: An example of a small Australian Food web.

One of the solution proposed was presented by Garlaschelli et al. [12] the idea is that the complexity of the network can be reduced according to two experimental facts. The first one is that the system is close. This means that every species ultimately must predate on primary producers that transform minerals, water and energy from sunlight into living matter. Secondly a layer division of the vertices according to the distance from the primary producers has an explicite meaning. Indeed at any level of predation only 10% of resources passes to the predator, this means that the largest the distance (in the topological sense of number of steps) from primary producers, the lower the amount of resources obtained.

Using these two empirical facts, the whole network can be described with a reasonable degree of accuracy by the spanning tree where cycles are removed by means of a BFS procedure starting on the primary producers. While the universality of the results obtained is still under debate [13], this method seems to well describe the various ecosystems.

- **Taxonomies**

Though modern systematic biologists consider it more reliable to describe relationships between species via phylogenetic trees (hierarchical structures following the steps of evolution from ancestral species to modern ones), also for their being less arbitrary, the classical taxonomic classification is still widely used. Introduced in the XVIII century by the Swedish naturalist Carl von Linné, taxonomy is based on morphological and physiological observations and it groups all species in a set of different hierarchical levels very much similar to a genealogic tree. Any classification group in the taxonomic tree is generally called a *taxon*. Because of the positive relationship between the phylogenetic relatedness of two taxa and their morphological and ecological similarity, the taxonomic tree of a particular flora should contain information on the processes that shape it. The distribution of the number of subtaxa per taxon at one specific taxonomic level has been widely studied starting with the work of Willis [15] and Yule [16]. Willis observed in 1922 that the distribution of the number n_g of genera containing a number n_s of species in the set of all flowering plants is a power-law with exponent about -1.5. In 1924 Yule proposed a branching process model to explain this distribution. Since then the shape of taxonomic abundance distributions has been object of study (later, in 1993, B. Burlando [17] extended these results to other pairs of taxonomic levels and observed similar behaviours for the distribution of the number n_t of taxa with n_{sb} subtaxa) [18,19,20] and models have been proposed in order to reproduce this behaviour [19,20].

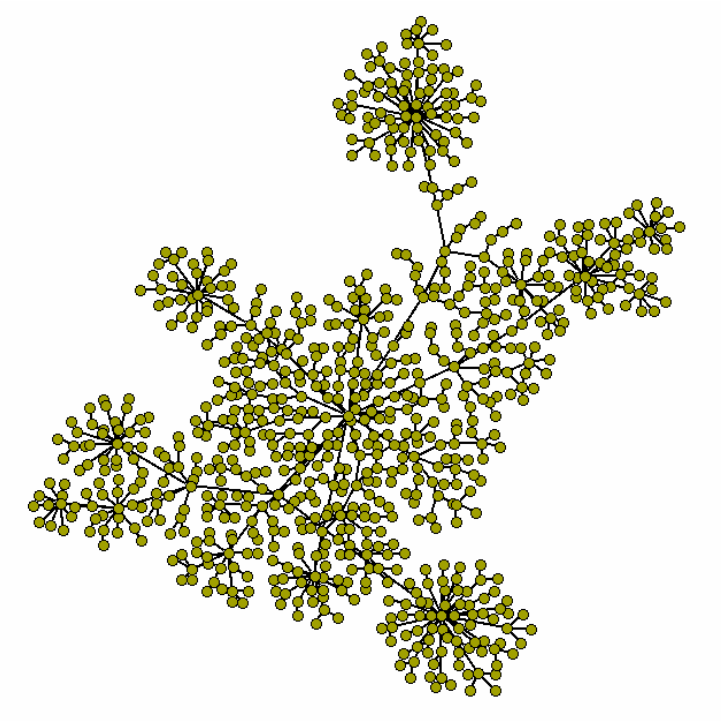


Figure 3: The taxonomic tree representing the Flora of Colosseo in 1874

Caretta Cartozo et al [21] show that for any of these data sets we find statistical properties that are stable in time and universal with respect to geographic and climatic environments. On this purpose we adopt a graph representation of the taxonomic tree. We assign a vertex for each taxon and an edge is drawn between two vertices *i,j* if the corresponding subtaxon *j* belongs to the taxon *i.* At the highest level, all species are eventually grouped in a same taxon; this implies that the resulting topology obtained is a particular kind of graph called a tree. In most cases the statistical properties of such graphs are universal (the frequency distribution of the number of links per site k (i.e. the degree) is distributed according to a power-law of the kind $P(k) \quad k^{-g}$ with an exponent g between 2 and 3) suggesting a possible unique mechanism for the onset of such common features [13].

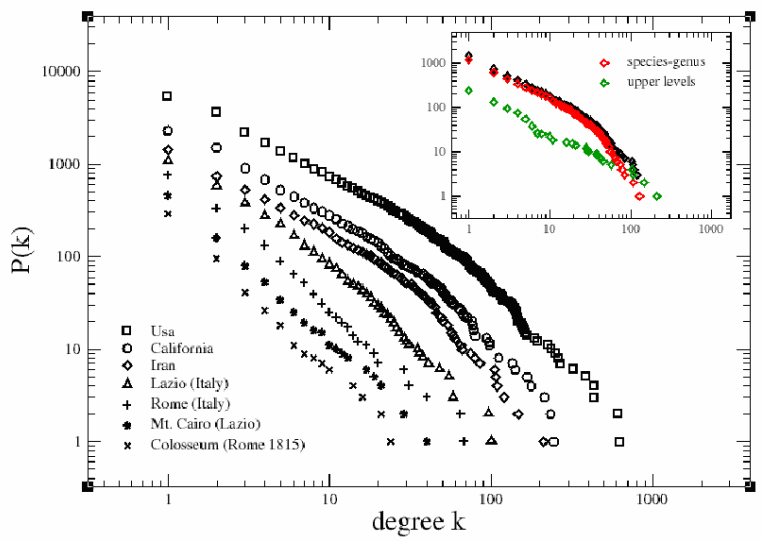


Figure 4: The degree distribution of various taxonomic trees

Even if at the moment it is impossible to determine what have been the evolutionary forces that shaped this peculiar distribution of offsprings between the various taxa, Caretta Cartozo et al. believe that they can use the value of the exponent in order to distinguish between real ecosystems and random collection of species. This result would allow to measure quantitatively the naturality of different environments.

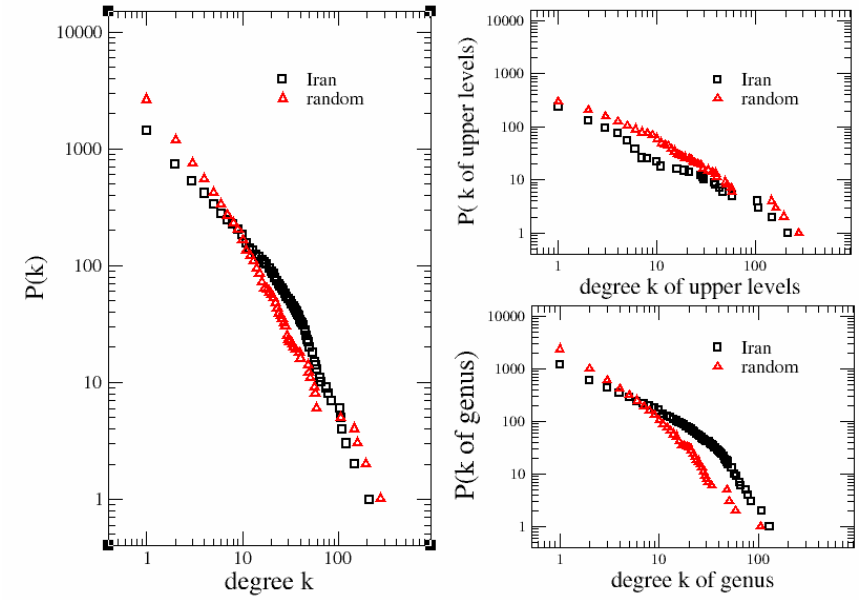


Figure 5: A comparison between real ecosystems and random collection of species.

12) Montoya, J.M.; Sole , R.V. ***Small world patterns in food webs.*** *Journal of Theoretical Biology*, **214**, 405—412 (2002).

13) • Garlaschelli D. Pietronero L.. Caldarelli G. Universal Scaling Relation in Food Webs, *Nature* **423,** 165 (2003).

14) • Arenas A. Camacho J. and Garlaschelli D. Pietronero L. Caldarelli G. Food Web Topology *Nature* **435** E3 and E4 (2005).

15) Willis, J. C. Age and area. Cambridge University Press, Cambridge 1922.

16) Yule, G. U. A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis, *Phil. Trans.* **B 213**, 21-87 (1924).

17) Burlando, B. The fractal dimension of taxonomic systems. *Journal of Theoretical Biology*, **146**, 99-114 (1990); *ibid*, The fractal geometry of evolution. *Journal of Theoretical Biology*, **163**, 161-172 (1993).

18) Enquist, B. J., Niklas, K. J. Invariant scaling relations across tree-dominated communities. *Nature*, **410,** 655-660, (2001).

19) Enquist, B. J., Haskell, J. P., Tiffney, B. H. General patterns of taxonomic and biomass partitioning in extant and fossil communities. *Nature*, **419**, 610-613, (2002).

20) Webb, C. O., Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. *American Naturalist* **156** 145-155 (2000)

21) • Caretta-Cartozo C., Barthelemy, M, Garlaschelli D., Ricotta C. Caldarelli G. Scale-free properties of taxonomy data. *Prolc. Nat. Acad. Sciences* **XX,** xxxx (2005).

Test Analysis for the WWW

One of the biggest challenges of text analysis techniques for the WWW is developing tools that are language independent. Many approaches to text analysis are based on preconstructed databases, e.g. WordNet [22]. In our work, we have tested the utility of spectral methods for clustering words according to their semantic and syntactic role in texts. Previous studies had shown that spectral methods can cluster words according by semantic field [23]. We have shown that spectral methods do cluster words according to their part-of-speech in a syntactic dependency network [24]. Although the network structure is given a priori by a syntactic dependency corpus, the network could be easily replaced by a word co-occurrence network that can be easily constructed from raw text [25, 26]. Fig. 6 shows a plot of the Romanian syntactic dependency network.

The extraction of information from texts needs knowing in advance which words within a sentence are more likely to be related. We have explored the probability that two words at a certain distance are syntactically related [27]. The analysis of the Euclidean distance between syntactically related words in a sentence suggested an exponential decay of the probability that two words are syntactically related with distance. A maximum entropy approach was used for explaining the exponential decay found in real sentences. The main message is that most of information within texts is at short distances (neighbours or second neighbours within a sentence). Long-distance syntactic relations are possible but rare. That theoretical insight suggests that applications can safely take advantage of the locality of syntactic links between words.

Many applications dealing with large texts, e.g. search engines, need some knowledge about the growth of vocabulary size (i.e. the number of different words). The growth of vocabulary size depends on the exponent of Zipf's law for word frequencies [28]. It is known that the exponent of Zipf's law deviates from its typical value in large texts [28, 29]. It was believed that those radical deviations were only found in large corpora. We have shown that those radical variations are also found in single author texts [30]. In the case of single author texts, the variations seem to be due to changes in the balance between maximizing the information transfer and saving cost in word use. The deviations in the exponent of Zipf's law in large corpora do not seem to be caused by that mechanism. Those should be the subject of further research.

A substantial amount of time was spent in improving previous models of network optimization in collaboration with L. Buriol. The main idea is trying to explain the topology of some scale free network (e.g. the WWW) through a process that minimizes the distance between links but minimizes the amount of links used [31]. The collaboration with L. Buriol consisted of improving the algorithms use by Ferrer i Cancho & Solé [31] so that larger networks could be studied and the topologies could be studied using the state of the art of network analysis tecniques. That project is still in progress.

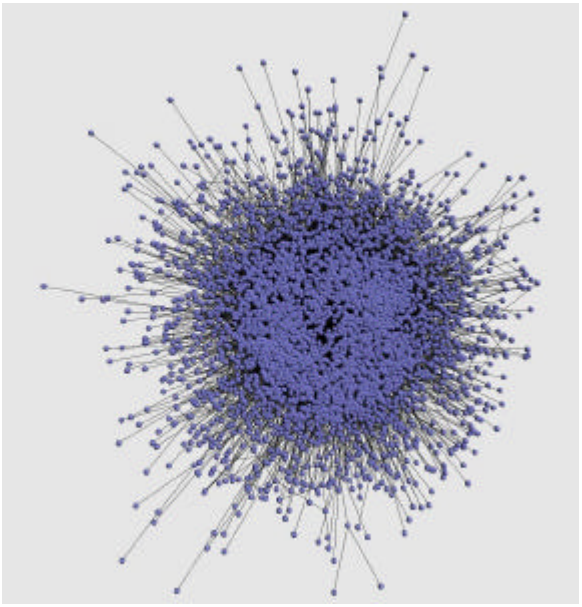


Fig. 6. The global syntactic dependency network of a Romanian corpus. (drawing by S. Valverde). Vertices are Romanian words and connections indicate syntactic dependencies. The network has 5,563 vertices.

22) Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the ACM 38, 39-41.

23) •Capocci, A., Servedio, V. D. P., Caldarelli, G. & Colaiori, F. (2005). Detecting communities in large networks. Physica A 350, 491-499.

24) •Ferrer i Cancho, R., Capocci, A. and Caldarelli, G. (2005). Spectral methods cluster words of the same class in a syntactic dependency network. cond-mat/0504165.

25) Ferrer i Cancho, R. Solé, R. V. (2001). The small-world of human language. Proceedings of the Royal Society of London B 268, 2261-2266

26) Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M. & Alon, U. (2004). Superfamilies of evolved and designed networks. Science 303, 1538-1542.

27) •Ferrer i Cancho, R. (2004). The Euclidean distance between syntactically linked words. Phys. Rev. E 70, 056135.

28) Ferrer i Cancho, R. and Ricard V. Solé (2001). Two regimes in the frequency of words and the origin of complex lexicons: Zipf's law revisited. Journal of Quantitative Linguistics 8, 165-173.

29) Montemurro, M. A. & Zanette, D. (2002). Frequency-rank distribution in large samples: phenomenology and models. Glottometrics 4, 87-98.

30) •Ferrer i Cancho, R. (2005). The variation of Zipf's law in human language. European Physical Journal B 44, 249-257.

31) Ferrer i Cancho, R. & Solé, R. V. (2003). Optimization in complex networks. In: Statistical Mechanics of complex networks, Pastor-Satorras, R. et al. (eds.). Lecture Notes in Physics 625, 114-125. Berlin: Springer.

GEOMETRIC SMALL WORLD NETWORKS

Small world networks are clear examples of systems that have to be studied as whole. They are graphs with a small diameter, consequence of a few long range connections (*shortcuts*), and a high degree of local interconnectedness: the global and local scales appear explicitly in the definition itself. Yet, they are not just mathematical structures: they are believed to play a pivotal role in the organization of social and communication networks, and there is growing evidence that they govern the large scale architecture of the brain. Petermann and De Los Rios [11] have provided a simple argument to discriminate between small-world and Euclidean networks as the distribution of the lengths of long range connections changes. According to their results, small-world networks occupy a wide region in the parameter space defining the shortcut distribution. Their findings also show that small-world networks can be realized in systems subject to strong resources constraints, and provide therefore a framework for the interpretation of data from real-world networks.

[32] T. Petermann and P. De Los Rios, *Spatial small-world networks: a wiring cost perspective,* arXiv:cond-mat/0501429.

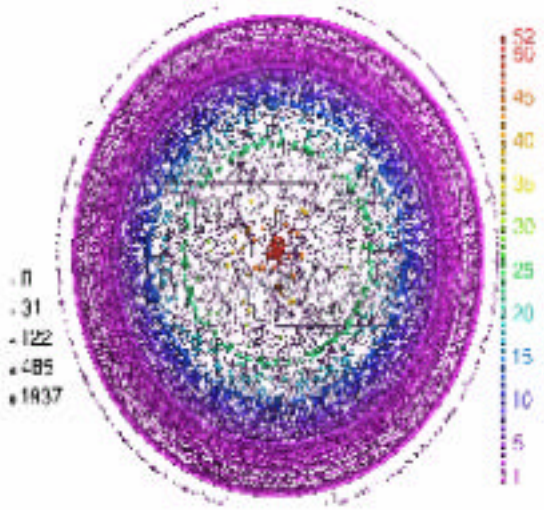
APPENDIX 1

1

STRUCTURE AND DYNAMICS OF COMPLEX NETWORKS

From Information Technology to Finance and Natural Science.

G. Caldarelli and A. Vespignani eds



Contents