Introduction to Complex Networks

Guido Caldarelli

INFM UdR ROMA1 Dip. Fisica, Università di Roma
“La Sapienza” P.le A. Moro 2 00185 Roma, Italy.

Abstract. We present here an introduction to the ideas and models that physicist developed in order to describe the graph or network structure in a variety of different systems. Firstly we give a very basic list of definition that can be of some help in approaching this field. After that we present a brief review of the state of art for the models.

INTRODUCTION

Recently there has been a great interest in the field of networks. As a matter of fact, dynamics and growth of networks shape today’s informational and social phenomena, ranging from the Internet and the WWW to economical networks and ecological food webs. The unifying feature of these networks is that their global structure and dynamic evolution are the result of locally interacting agents distributed in the network. Such systems are therefore a paramount example of complex systems whose presence is becoming more and more evident in physics, biology, and mathematics as well as in computer science. From a mathematical point of view, complex networks are sets of many interacting components whose elaborate collective behaviour cannot be directly predicted and characterised in terms of the behaviour of their individual components. When the interactions between single components are suitably modelled, components can describe many different real-world units such as internet providers, electricity providers, economical agents, ecological species, etc. The dynamics of the whole system can describe the emerging global behaviour such as the internet traffic, electricity supply service, market trend, environmental resources depletion etc. As for a non-rigorous characterisation of this complexity at least in Physics, we refer to the class of phenomena, like deposition, corrosion, cracking, growth of colonies and in general all the phenomena where the simple basic interactions between agents are such that to produce self-similar structures. These self-similar (i.e. fractal. That is they show the same shape at any level) structures are rather peculiar since they show correlation in shape at large distances both in space and in time *a priori* this correlation is unpredictable from the microscopic dynamics. Growing networks presents all these properties. They effectively start from a small collection of components. During growth they start to develop some nontrivial features. The typical signatures of such features are the power-law distributions in the quantities of interest. Currently different models and theories have been proposed, but in order to assess their validity a remarkable set of data should be collected. Growing networks then represent at this time an exciting research field. We hope to present here the current state of the art with the many open problems and the applications to come.

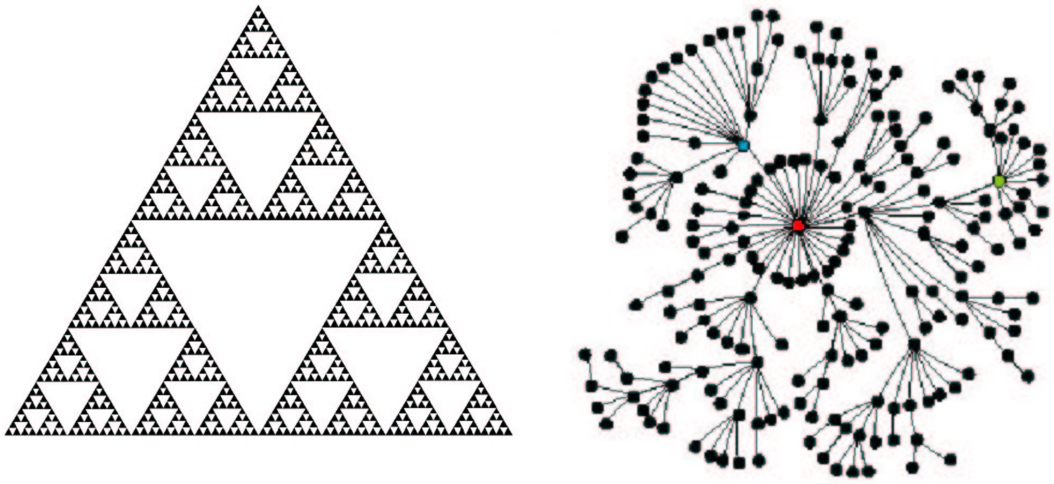

FIGURE 1. On the left a typical fractal, on the right a complex network. IN the latter case the self-similarity is related to the topology.

BASIC DEFINITIONS IN GRAPH THEORY

Networks can be represented by graphs, where by graphs we mean the mathematical entities introduced by the seminal work of Erdős and Renyi[1, 2, 3]. Graphs are composed by sites connected by links. Everywhere we shall call the sites as vertices and the links as edges. Edges can have arrows or not. In the first case the graph is oriented. We shall indicate in the following a graph composed by v vertices and e edges as $G(v,e)$. In order to compute some quantities of interest for our cases of study let us introduce some basic definitions. For a deeper discussion on the mathematical basis of Graph Theory we suggest to consult Ref.[4].

- The graph **order** is the number of its vertices.
- The graph **size** is the number of its edges.
- The **degree** of a vertex in a graph is the number of edges that connects it to other vertices.
- In the case of an oriented graph the degree can be distinguished in **in-degree** and **out-degree**.
- Whenever all the vertices share the same degree the graph is called **regular**.
- A series of consecutive edges forms a **path**. When end-vertices coincide the path is a **circuit**. If the other vertices and all the edges are all distinct the circuit is a **cycle**.
- A graph is connected if a path exists for any couple of vertices in the graph.
- A graph with no cycles is a **forest**. A **tree** is a connected forest.
- The **distance** between two vertices is the shortest number of edges one needs to travel to get from one vertex to the other.
- Therefore the neighbours of a vertex are all the vertices which are connected to that vertex by a single edge.

- The **diameter D** of a graph is the longest distance you can find between two vertices in the graph.
- A complete bipartite **clique** Ki,j is a graph where every one of i nodes has an edge directed to each of the j nodes.
- A **bipartite core** Ci,j is a graph on i+j nodes that contains at least one Ki,j as a subgraph.
- A complete bipartite **clique** Ki,j is a graph where every one of i nodes has an edge directed to each of the j nodes.
- A **bipartite core** Ci,j is a graph on i+j nodes that contains at least one Ki,j as a subgraph.
- The **clustering coefficient** C is a rougher characterization of clustering with respect to the clique distribution. C is given by the average fraction of pair of neighbours of a node that are also neighbours each other. For an empty graph En C=0 everywhere. For a complete graph Kn, C=1 everywhere.
- The **betweenness** b_i of a vertex gives the number of different distances between all the couple of vertices in the graph that pass through vertex i.

This very schematic list is only meant to provide some basic notions that could help in recognizing some typical patterns of graph in the physical world. Consider for example a scientist that is interested in designing a functional way to share information. It would be sensible in this case to look at the WWW. The billions of pages of the WWW can be considered as vertices, and their hyperlinks as edges of a giant graph. The out-degree of a site is then the number of hyperlinks present in it; more importantly in order to determine the success of a web-page, one can consider the number of other pages that point to it. They of course give the in-degree of the page. A cyber-community of football fans are likely to form a bipartite core in this graph, whilst cycles allow us to reach a site even if our favourite portal pointing to it is temporarily out of order.

In all these examples, the degree k of a vertex, *i.e.* the number of arcs linking it to other vertices, is power-law distributed, $P(k) \sim k^{-\gamma}$. SF networks also present the, so called, *small-world* phenomenon [5], that is, by few selected jumps (that can be either short- and long range) it is possible to reach very different regions of the system and apparently distant environments.

ectionModels

the Erdős Rényi model

In this model, the number of vertices is fixed and the probabilty to draw a link between any couple of vertices is constant. Therefore the probability to have a vertex whose degree is k is simply

$$\frac{N!}{k!(N-k)!}\, p^k(1-p)^{N-k} \tag{1}$$

In the limit of a large N this tends to a Poissonian distribution peaked around a typical value. It has to be noticed that many data instead show scale free behaviour. Despite of the name of Random Graph model, this model seem to reproduce the properties of some peculiar networks.

the Barabàsi Albert model

To understand how SF networks arise, Barabási and Albert[6] introduced the concepts of *growing networks* and of *preferential attachment*. In their original (BA) model, networks grow at a constant rate (modeling the fact that new Web pages are continuously created, new proteins emerge by mutation, and so forth). New vertices are attached to older ones with a probability $\pi(k)$ which is a (linearly [7]) growing function of the number of preexisting links, k, at every site. In this way, a highly connected vertex is more likely to receive further links from newly arriving vertices: this is the so-called "rich get richer" rule. The probability distribution for the degree in the simple case of one edge added at time is given by $P(k) \propto k^{-3}$.

The Copying models

In some other, recently proposed, models of protein interaction networks [8, 9] and of the WWW [10], new vertices (proteins and Web pages respectively) are added by copying (replicating) existing vertices, borrowing some of their links and adding some new others. The model is related to the supposed genesis of an html document. Since it is very unlikely to start from scratch in the creation of the own homepage, one can think to borrow the best friend home page and change only a little bit of the content. In this self-consistent way one can take into account the formation of cyber communities. It has been shown that this mechanism leads also to an *effective* preferential attachment rule.

Generalized Erdős-Renyi model

Yet, although in some contexts preferential attachment can be a very reasonable assumption, in many others it is certainly not. In particular, in some situations, the information about the degree of each and every single vertex is not available to newly added sites, neither in a direct nor in an effective way. Instead, it is reasonable that two vertices are connected when the link creates a mutual benefit (here we restrict ourselves to bidirectional links) depending on some of their intrinsic properties (authoritativeness, friendship, social success, scientific relevance, interaction strength, etc). Therefore, it is reasonable to expect that for some of these systems the $P(k)$ scale free behaviour (when existing) has an origin unrelated to preferential attachment.

In order to explore this simple idea, another network-building algorithm is possible:

- start by creating a total (large) number N of vertices. At every vertex i a fitness x_i, which is a real number measuring its importance or rank, is assigned. Fitnesses are random numbers taken from a given probability distribution $\rho(x)$.

- For every couple of vertices, i, j, a link is drawn with a probability $f(x_i, x_j)$ (f a symmetric function of its arguments) depending on the "importance" of both vertices, *i.e.* on x_i, x_j.

A general expression for $P(k)$ can be easily derived. Indeed, the mean degree of a vertex of fitness x is simply

$$k(x) = N \int_0^\infty f(x,y)\rho(y)dy = NF(x) \tag{2}$$

(with $x_i \in (0,\infty)$). Assuming $F(x)$ to be a monotonous function of x, and for large enough N, we have the simple relation

$$P(k) = \rho\left[F^{-1}\left(\frac{k}{N}\right)\right]\frac{d}{dk}F^{-1}\left(\frac{k}{N}\right). \tag{3}$$

For finite values of N corrections to this equation emerge [11]. As a particular example, consider $f(x_i, x_j) = (x_i x_j)/x_M^2$ where x_M is the largest value of x in the network. Then

$$k(x) = \frac{Nx}{x_M^2}\int_0^\infty y\rho(y)dy = N\frac{<x>x}{x_M^2} \tag{4}$$

and we have the simple relation

$$P(k) = \frac{x_M^2}{N<x>}\rho\left(\frac{x_M^2}{N<x>}k\right). \tag{5}$$

A particularly simple realization of the model emerges if we consider power-law distributed fitnesses. This choice can be naturally justified by arguing that power-laws appear rather generically in many contexts when one ranks, for example, people according to their incomes or cities according to their population, etc. This is the so-called Zipf law which establishes that the rank $R(x)$ behaves as $R(x) \propto x^{-\alpha}$ in a quite universal fashion [12]. The reason for the ubiquitous presence of the Zipf law yields on the multiplicative nature of the intrinsic fluctuations which generically leads to flat distributions in logarithmic space and, consequently, to power-laws [12].

Clearly, if $\rho(x) \sim x^{-\beta}$ (Zipf's behaviour, with $\beta = 1 + 1/\alpha$ [12]) then, using eq.(5), also the degree distribution $P(k)$ is a power-law and the network shows SF behaviour.

This result is hardly surprising: from SF fitnesses we generate SF networks, but still it provides a new generic path to SF networks and takes into account the widespread occurrence of the Zipf's behaviour in nature. In order to extend this result and check whether SF networks can be generated even when $\rho(x)$ is not SF itself, we consider an exponential distribution of fitnesses, $\rho(x) = e^{-x}$ (representing a random, Poisson distribution) and $f(x_i, x_j) = \theta(x_i + x_j - z)$, where $\theta(x)$ is the usual Heaviside step function. This represents processes where two vertices are linked only if the sum of their fitnesses is larger than a given *threshold* z. Using these rules we obtain analytically (and confirm in computer simulations) that $P(k) \sim k^{-2}$. This leads to the non-trivial result that *even non scale-free fitness distributions can generate scale-free networks*. Also different implementations of the threshold rule, such as $f(x_i,x_j) = \theta(x_i^n + x_j^n - z^n)$ (where n is an integer number) give rise to the same inverse square behavior (although, in some cases, with logarithmic corrections).

Let us stress that the model, as defined, has a diverging average connectivity in the large N limit, as can be easily inferred from Eq.(2); *i.e.* it is severely *accelerated* [13]. Nevertheless we can introduce in a rather natural way an upper cut-off accounting for the fact that every site has a limited information on the rest of the world and, therefore, connection is attempted with a finite number, m, of different sites. Alternatively, vertices can be linked with the above rule and, after that, links are kept with probability p (so that, for example, $pN = m$). By including this modification, the N factor in Eq.(2), is substituted by m, and the connectivity is finite in the thermodynamic limit. In order to generate different accelerated networks (with the averaged connectivity not reaching a stationary value but growing with N in different possible ways [13]) other selection rules can be easily implemented.

Computer simulations of this model show that networks with power-law distributed fitnesses, and different values of β, show nearly constant $k_{nn}(k)$'s and $c(k)$'s, just as occurs for the original BA model [6]. The distribution of betweenness decays as a power law with an exponent $\gamma_b \approx 2.2$ for $\gamma = 2.5$ and $\gamma = 3$, and $\gamma_b \approx 2.6$ for $\gamma = 4$. This is in good agreement with what conjectured in Ref. [14]: all networks with $3 \geq \gamma > 2$ can be classified in only two groups according to the value of γ_b ($\gamma_b = 2$ and $\gamma_b = 2.2$, respectively), while for larger values of γ, larger non-universal values of γ_b are reported.

The exponential case behaves in a different way: for a network of size $N = 10^4$, $z = 10$, and $m = N$ we find $< d >= 2$, $< c >\simeq 0.1$ and $< b > /N \simeq 0.1$, but a power-law behavior is found for the clustering magnitudes, *i.e.* $< k_{nn} >\propto k^{-0.85}$ and $c(k) \propto k^{-1.6}$. The betweenness distribution instead, shows an unexpected behavior, giving a power-law tail with an exponent $\gamma_b \approx 1.45$. It is worth remarking that our model having $\gamma = 2$ is not included in the previously discussed classification of betweenness exponents [14].

In summary, we have presented an alternative model to justify the ubiquity of SF networks in nature. It is a natural generalization of the standard Erdős-Rényi. The main result is that emergence of SF properties is not necessarily linked to the ingredients of growth and preferential attachment. Instead, static structures characterized by quenched disorder (for different disorder distributions) and threshold phenomena, may generate effects very similar to those measured in the real data.

ACKNOWLEDGMENTS

We thank FET Open Project IST-2001-33555 COSIN `http://www.cosin.org` for support

REFERENCES

1. P. Erdős and A. Rényi, *Publ. Math. Debrecen* **6**, 290 (1959).
2. P. Erdős and A. Rényi, *Publ. Math. Inst. Hung. Acad. Sci.* **5**, 17 (1960).
3. P. Erdős and A. Rényi, Bull. Inst. Int. Stat. **38**, 343 (1961).

4. B. Bollobás, *Random Graphs* (Academic Press, London) (1985).
5. D. Watts, *Small-Worlds*, Princeton Univ. Press (1999).
6. R. Albert and A.-L. Barabási, *Rev. Mod. Phys.* **74**, 47 (2002), and references therein.
7. SF networks are generated only if $p(k) \sim k$; sub-linear $p(k)$'s produce skewed but not SF networks; P. L. Krapivsky and S. Redner, *Phys. Rev. E*, **63**, 066123 (2001).
8. A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani, *cond-mat/0108043*.
9. R. V. Solé, R. Pastor-Satorras, E. D. Smith, and T. Kepler, *Adv. Compl. Systems* **5**, 42 (2002).
10. R. Kumar, P. Raghavan, S. Rajalopagan, and A. Tomkins, Proc. of the 9th ACM Symposium on Principles of Database Systems, 1 (1999).
11. P. L. Krapivsky and S. Redner, *cond-mat/0207107*.
12. M. Marsili and Y-C. Zhang, *Phys. Rev. Lett.* **80**, 2741 (1998). R. Cont and D. Sornette, J. Phys. I **7**, 431 (1997).
13. S. N. Dorogovtsev and J. F. F. Mendes, in "*Handbook of Graphs and Networks: from the Genome to the Internet*" eds. S. Bornholdt and H. G. Schuster, pag. 320 (Wiley, Berlin), 2002. *cond-mat/0204102*.
14. K. -I. Goh, B. Kahng, and D. Kim, *Phys. Rev. Lett.* **87**, 278701 (2001). K. -I. Goh, et al. *cond-mat/0205232*.