Dynamic exploration of networks: from general principles to the traceroute process

L. Dall'Asta[a]

Abdus Salam International Center for Theoretical Physics, Strada Costiera 11, 34014, Trieste, Italy

Abstract. Dynamical processes taking place on real networks define on them evolving subnetworks whose topology is not necessarily the same of the underlying one. We investigate the problem of determining the emerging degree distribution, focusing on a class of tree-like processes, such as those used to explore the Internet's topology. A general theory based on mean-field arguments is proposed, both for single-source and multiple-source cases, and applied to the specific example of the traceroute exploration of networks. Our results provide a qualitative improvement in the understanding of dynamical sampling and of the interplay between dynamics and topology in large networks like the Internet.

PACS. 89.75.-k Complex systems – 89.20.Hh World Wide Web, Internet – 89.75.Fb Structures and organization in complex systems

1 Introduction

The theory of *complex networks* is a primary tool in the interdisciplinary research in complex systems. It provides a simple (coarse-grained) description of natural as well as social and technological complex systems, by means of mapping their elements, and the interactions among them, on the vertices and the edges of a graph [1]. Different systems correspond to microscopically different networks, but their general statistical properties may display similarities. Using the tools of statistical physics in analyzing experimental data, researchers have obtained important insights on the structure of these complex networks, discovering universal properties that are common to very different systems. In particular, most of real networks seem to be sparse, with a very small diameter (small-world property), and to present large fluctuations in the connectivity (degree) of the nodes.

The theoretical models claiming to explain the origins of these universal properties can be roughly divided in two classes: static models, in which the whole network structure is assigned a priori, and growing models, in which new nodes attach to a pre-existing network by means of a given number of links [2–5]. Much less attention was instead devoted to the other very important issue concerning the relation between networks structure and the processes used to collect the experimental data. Some questions arise naturally: are the experimental data really reliable? can the ubiquity of some properties be the result of systematic biases introduced by the sampling process itself?

These questions have been recently addressed in a series of works [6–10] that study the sampling of networks from both static and dynamic points of view. Among them, the work by Petermann and De Los Rios [6] is the first in which the idea that different generation algorithms may lead to different topological properties is exploited to show analytically that the observed power-law degree distributions can be an artifact of biases affecting the sampling techniques.

This issue is particularly relevant for the physical Internet, that is the prototype of a complex network: a huge number of computers are connected together to form a continuously evolving, self-organizing network, on which a large variety of dynamical processes take place [11]. As the robustness and performances of large technological networks strongly depend on their topology, an accurate, unbiased knowledge of the Internet is necessary to support the work of practitioners in improving its structure and functioning.

Motivated by the importance of correctly knowing network's topology, the present work is devoted to clarify the mechanisms governing the main techniques used in the dynamical sampling of complex networks. In order to accomplish this task, we first present the general framework that applies to a large class of tree-like dynamical processes, then we focus on the *traceroute sampling* of networks with homogeneous and heterogeneous degree profiles. We show that the role played by the dynamics is very different in homogeneous and heterogeneous networks. In the latter case, that should correspond to the case of the Internet, the sampling process introduces some level of

---

[a] e-mail: dallasta@ictp.it

inhomogeneity that paradoxically increases the sampling accuracy of the tail of the degree distribution.

Our results are in agreement with previous analytical and numerical achievements, but allow a more unitary view of the problem. Dynamical sampling techniques are examples of the interplay between networks topology and the dynamics occurring on it. Therefore the dynamical approach presented in this paper could be considered as a theoretical improvement with respect to previous analytical studies based on nodes exposure methods [12,13], that are essentially static algorithms.

Moreover, our method can be easily adapted to study other sampling algorithms expressly designed to acquire ego-centric views of a network, such as network's tomography [14,15] and crawling algorithms in community networks (the WWW, P2P networks, etc.) [16]. Similar dynamical patterns are observed also in network processes that are not directly related to sampling, such as epidemics [17], rumors spreading [18], diffusion of innovations [19] and threshold models [20].

In summary, we provide a unified description of a class of dynamical phenomena in terms of tree-like evolving structures on networks, and explain how the dynamical rules influence the emerging topological properties, by focusing on the degree distribution.

The paper is organized as follows. In Section 2, we present the general theoretical formalism that can be used to study the degree distribution generated by tree-like processes on networks. The important application to the traceroute problem is discussed in Section 3. Some conclusions are presented in Section 4 together with examples of other possible applications.

2 General formalism for tree-like processes

Among the various types of algorithms and dynamical models evolving on networks, we take into account those corresponding to the following general dynamical picture, that is also valid for the special case of traceroute-like explorations. Let us assume the process starts from a single node and propagates iteratively throughout the network. At each temporal step, some nodes at the interface of the growing cluster are selected and some of their still unreached neighbors are visited. The latter ones become part of the interface, while the former interfacial nodes are moved to the bulk of the cluster (a sketch of the dynamics is reported in Fig. 1). During the dynamics we can always identify three distinct classes of nodes: bulk, interfacial, and unreached nodes. When the above process takes place on a random network, in the limit of large network's size N, the overall dynamics is well represented by the temporal evolution of some mean-field quantities, the densities of bulk nodes $b(t)$, interfacial nodes $i(t)$, and unreached nodes $u(t)$. Obviously, $u(t) + i(t) + b(t) = 1$ always during the dynamics.

On a generic random network, however, the degree is not fixed, the nodes being divided in degree classes. The global densities are replaced by degree-dependent
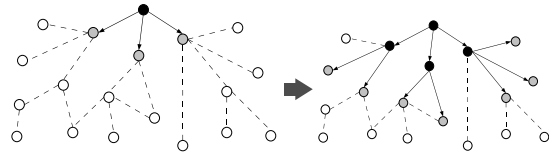
Fig. 1. Sketch representing the evolution of the considered class of dynamical processes. Starting from a single source, the nodes of the network are progressively visited. Once the nodes at the interface (grey nodes) have spread the process towards their still unknown neighbors (white nodes), they are moved to the bulk (black nodes).

partial densities $b_k(t)$, $i_k(t)$, and $u_k(t)$. The partial density for unreached nodes of degree k is defined as the fraction of nodes of degree k that are still unreached at time t. The normalization relation is $\sum_k P(k)u_k(t) = u(t)$, where $P(k)$ is the degree distribution of the underlying network. The other quantities are defined similarly.

More in general, one may be interested in situations in which the network is still maximally random (with degree distribution $P(k)$) but with some degree correlations, expressed by the conditional probability $P(k|h)$ that a node of degree k is linked to a node of degree h. One can also consider the nodes divided in *types* α, i.e. discrete or continuous states defined on the nodes. Types are used to create multi-partite networks and to encode some non-topological feature. Another way to account for non-topological properties is that of putting weights on the links, that depend only on the degrees and the types of the extremities. Link weights may account for dynamical properties of the process, such as the transmissibility of a disease [21]. The internal structure of the population is taken into account defining degree-dependent and type-dependent partial densities of $u_k^\alpha(t)$, $i_k^\alpha(t)$, and $b_k^\alpha(t)$. The global densities are recovered by averaging over all distributions, i.e. $u(t) = \sum_\alpha \mathcal{P}(\alpha) \sum_k P(k)u_k^\alpha(t)$. In order to simplify the formalism, here on we limit our analysis to single-type undirected random markovian networks [22], that are maximally random graphs completely defined by the degree distribution $P(k)$ and the degree correlations $P(k|h)$.

According to this mean-field approximated description of the processes, the temporal evolution of the partial densities satisfies a system of differential equations of the type,

$$
\begin{cases}
\frac{d}{dt}u_k(t) = f_u(\{u_h(t)\}, \{i_h(t)\}, t, \cdots) \\
\frac{d}{dt}i_k(t) = f_i(\{u_h(t)\}, \{i_h(t)\}, t, \cdots) \\
\frac{d}{dt}b_k(t) = f_b(\{i_h(t)\}, t, \cdots),
\end{cases}
\tag{1}
$$

where the arguments of the functions $f_x(\cdot)$ depend on the general form of the dynamics described above. For instance, f_b is not expected to depend on unreached nodes, f_u on the bulk nodes, etc. The system is generally coupled and non-linear, and admits an explicit solution only in very special cases.

In statistical physics and theoretical biology, continuous mean-field dynamical equations are commonly used

to study models of population dynamics. The dynamical picture emerging from these mean-field models allows to understand the qualitative behavior of complex phenomena occurring in real systems. At the same time, it is worth noting that the rigorous derivation of differential equations for random processes on random graphs has been introduced in the mathematic literature only recently, by Wormald [23], and then applied to several problems, including algorithms for the generation of random graphs with a given degree sequence [24] and random k-SAT problems [25]. Wormald's differential equations method provides a powerful tool to prove rigorous bounds for interesting quantities (e.g. via distribution's moments) in discrete-time combinatorial processes. Here we limit our analysis to a qualitative topological characterization of the emerging degree distribution obtained within a purely mean-field statistical physics approach. However, a rigorous formalization of the present approach is desirable as well.

2.1 Single-source processes

The degree distribution $\tilde{P}(k)$ of a subnetwork is related to the degree distribution $P(k)$ of the underlying one by

$$\tilde{P}(k) = \sum_{\ell=k}^{\infty} P(\ell) Q(k|\ell) \qquad (2)$$

where $P(\ell)$ is the degree distribution, that defines the probability of picking up a node of degree ℓ in the original network, and $Q(k|\ell)$ is the conditional probability of observing a node of degree k in the subnetwork if its real degree in the complete network is ℓ. In a dynamical framework, the sampling probability depends on the temporal evolution of the overall process. At the beginning, the neighborhood of interfacial nodes is mainly composed of unreached ones, whereas in the final stage of the dynamics most of the nodes have already been visited. The probabilities $P(\ell)$ and $Q(k|\ell)$ are thus replaced by time-dependent quantities $P_t(\ell)$ and $Q_t(k|\ell)$, that are defined by the evolution rule of the dynamical processes itself.

Using the properties of the dynamics, $P_t(\ell)$ can be expressed as the probability of picking up a node of degree ℓ among the interfacial nodes of the growing cluster at a time t, i.e. $P_t(\ell) = P(\ell)i_\ell(t)/i(t)$. Then, the way in which the neighbors of this node are selected and visited depends strictly on the details of the dynamical model. When the growing cluster has a tree-like structure, a node of degree ℓ has only one incoming edge, and the remaining $\ell - 1$ links are used to propagate towards unreached neighbors. Let us call ρ_h the probability of visiting a neighbor of degree h of an interfacial node of degree ℓ (it is a function of $u_h(t)$), the conditional probability $Q_t(k|\ell)$ becomes

$$Q_t(k|\ell) = \binom{\ell-1}{k-1} \left[\sum_h P(h|\ell)\rho_h[u_h(t)] \right]^{k-1}$$
$$\times \left[1 - \sum_h P(h|\ell)\rho_h[u_h(t)] \right]^{\ell-k} . \qquad (3)$$

Putting together these two terms and recalling that the global topology is given by averaging over the whole temporal spectrum, we obtain the following expression for the degree distribution $\tilde{P}_1(k)$ of the tree-like structure emerging from the dynamics,

$$\tilde{P}_1(k) = \sum_{\ell=k}^{\infty} \tilde{P}_1(k,\ell) = \sum_{\ell=k}^{\infty} \frac{1}{T} \int_0^T P_t(\ell) Q_t(k|\ell) \; dt$$
$$= \sum_{\ell=k}^{\infty} \frac{1}{T} \int_0^T \frac{P(\ell)i_\ell(t)}{i(t)}$$
$$\times \binom{\ell-1}{k-1} \left[\sum_h P(h|\ell)\rho_h[u_h(t)] \right]^{k-1}$$
$$\times \left[1 - \sum_h P(h|\ell)\rho_h[u_h(t)] \right]^{\ell-k} \; dt, \qquad (4)$$

where T is the maximum sampling time. In the above equation we have also introduced the joint degree distribution $\tilde{P}_1(k,\ell)$ (of observing a node of degree k with real degree ℓ), that will be useful in the following.

The generalizations to multi-type and weighted networks are straightforward once one has correctly considered the evolution equations for the partial densities of bulk, interfacial and unreached nodes. Again, the use of the Bernoulli sampling technique (binomial probability) in selecting neighboring nodes is justified by the spreading like character of the dynamics considered here. For different classes of dynamical processes, e.g. threshold processes [20], the selection mechanism should be modified.

2.2 Multi-source processes

It frequently happens that several processes are running at the same time, so that the overall measure is obtained merging several single-source structures. This is indeed the case of Internet's mapping projects, in which single (spanning) trees from different source nodes are merged together in order to get more accurate mappings of the underlying topology. A completely different example is provided by the overall infection profile in a population that is object of multiple non-interacting disease spreadings (e.g. e-viruses and worms in P2P communities), each one generating a sort of causal tree.

In order to study multi-source processes, and the emerging degree distribution, we put forward an approximated method based on a simple mean-field argument for the overlap probability of uncorrelated trees. A direct generalization of the single source method presented above is somewhat tricky, since in any node one edge per process is used as an incoming edge and is not available for spreading. Increasing the number of sources one should take care of all possible combinations of these incoming links, that becomes rapidly very complicated. Nonetheless, this approach leads to some valuable approximation. Let us consider a process with two sources, and a node of degree $\ell + 1$ for which we fix the incoming edge, so that

both trees reach the node from such an edge. Moreover, we assume that the two sampling processes are completely independent and uncorrelated. With these hypotheses, the observed degree distribution, obtained merging two trees is

$$
\begin{aligned}
\tilde{P}_2(k+1) = \sum_{\ell=k}^{\infty} P(\ell+1) \frac{1}{T^2} \int_0^T dt_1 dt_2 \\
\times \frac{i_{\ell+1}(t_1)}{i(t_1)} \frac{i_{\ell+1}(t_2)}{i(t_2)} \sum_{m,n=0}^{\ell} \sum_{r=0}^{\ell} \binom{\ell}{m} \binom{\ell-m}{n-r} \\
\times \binom{m}{r} \delta(k-m-n+r)[\bar{u}(t_1)\bar{u}(t_2)]^r \\
\times [\bar{u}(t_1)(1-\bar{u}(t_2))]^{m-r} [\bar{u}(t_2)(1-\bar{u}(t_1))]^{n-r} \\
\times [(1-\bar{u}(t_1))(1-\bar{u}(t_2))]^{\ell-m-n+r},
\end{aligned} \quad (5)
$$

where $\bar{u}(t) = \sum_h P(h|\ell)\rho_h[u_h(t)]$ and $\delta(x)$ is the Kronecker's symbol. Reordering the terms in equation (5), and using the result for single-source processes, we get

$$
\begin{aligned}
\tilde{P}_2(k+1) = \sum_{\ell=k}^{\infty} P(\ell+1) \\
\times \sum_{m,n,r=0}^{\ell} \delta(k-m-n+r)\mathcal{B}(\ell,m,n,r) \\
\times \frac{\tilde{P}_1(m+1,\ell+1)}{P(\ell+1)} \frac{\tilde{P}_1(n+1,\ell+1)}{P(\ell+1)}
\end{aligned} \quad (6)
$$

in which $\mathcal{B}(\ell,m,n,r)$ is the hypergeometric distribution

$$
\mathcal{B}(\ell,m,n,r) = \binom{\ell-m}{n-r}\binom{m}{r}/\binom{\ell}{n}.
$$

The case in which both trees reach a node through the same edge is obviously very special, as well as equation (5) that does not hold in general. However, one can exploit the picture emerging from equation (6) and generalize it to be valid whatever the choice of the incoming edges. Unlabeling the incoming edges and considering them like the other edges discovered during single-source processes, we get the following approximation for the observed degree distribution in a process with two sources,

$$
\begin{aligned}
\tilde{P}_2(k,\ell) \approx P(\ell) \sum_{m,n=1}^{\ell} \sum_{r=0}^{\ell} \mathcal{B}(\ell,m,n,r) \\
\times \frac{\tilde{P}_1(n,\ell)}{P(\ell)} \frac{\tilde{P}_1(m,\ell)}{P(\ell)} \delta(k-n-m+r),
\end{aligned} \quad (7)
$$

the sums over m and n start from 1 since we assume that all nodes are discovered (the minimum observed degree being 1). Note that even if two consecutive processes are dynamically uncorrelated, the topological and functional properties of the underlying system always introduce some correlations. For instance, in real networks there are very central nodes that bear a large fraction of the traffic, the so-called *backbone nodes*. Depending on the process it may

be easier or more difficult to traverse these nodes. Again, in the Internet there are administrative policies governing local routing systems, therefore in some cases it is impossible to visit the entire neighborhood of a node. This kind of correlations cannot be easily included in the above mean-field analysis. On the other hand, Internet's local correlations may rapidly change in time, as they are affected by traffic congestions and routers' failures; therefore the average qualitative behavior of real processes should be close to the uncorrelated one [26].

Writing $\tilde{R}_2(k,\ell) = \tilde{P}_2(k,\ell)/P(\ell)$, equation (7) can be rewritten

$$
\begin{aligned}
\tilde{R}_2(k,\ell) = \sum_{m,n=1}^{\ell} \sum_{r=0}^{\ell} \delta(k-n-m+r) \\
\times \mathcal{B}(\ell,m,n,r)\tilde{R}_1(n,\ell)\tilde{R}_1(m,\ell),
\end{aligned} \quad (8)
$$

and the merging process can be easily generalized to any number ν of sources exploiting the recursion relation,

$$
\begin{aligned}
\tilde{R}_\nu(k,\ell) = \sum_{m,n=1}^{\ell} \sum_{r=0}^{\ell} \delta(k-n-m+r) \\
\times \mathcal{B}(\ell,m,n,r)\tilde{R}_1(n,\ell)\tilde{R}_{\nu-1}(m,\ell).
\end{aligned} \quad (9)
$$

The degree distribution of the network obtained merging ν trees is then given by $\tilde{P}_\nu(k) = \sum_\ell P(\ell)\tilde{R}_\nu(k,\ell)$. Note that equation (9) is general and holds, within the validity of the approximation, for any type of process in the class under study, while the explicit expression of $\tilde{R}_1(k,\ell)$ as well as the correct form of $\mathcal{B}(\ell,m,n,r)$ depend on the details of the dynamics.

In the next section, we show how these methods can be applied to the traceroute model that describes the experiments used to determine the topology of the Internet.

3 Application to the internet's mapping techniques

The Internet's topology can be studied at different levels. The most detailed Internet's descriptions are obtained at the level of single *routers*, but coarse grained representations are usually preferred for the possibility of obtaining a more reliable picture of the system (see Ref. [11] for a simple introduction to the networked representations of the Internet). Routers sharing the same administrative policies are divided into *Autonomous Systems* (AS), that define the most important coarse-grained level of the Internet's topology. The first maps of the Internet were collected mainly at the AS level, using empirical data extracted from BGP tables together with those obtained by dynamical sampling methods based on traceroute measurements from single source [27–29]. According to these maps, that collect partial views of the net from some favored points, the Internet should be a very heterogeneous network with approximately power-law degree distribution $P(k) \propto k^{-\gamma}$, and $\gamma \simeq 2.1 \div 2.4$ [28,29]. Because of

the peculiar properties of scale-free networks, this discovery motivated a series of theoretical works in which toy-models of the Internet's mapping process were proposed and analysed in order to justify or question this empirical evidence [8,10,30,31].

In a traceroute experiment, a given number of probes are sent from a source to a set of target nodes, tracing back the path followed during the exploration. These probes are data packets that follow the same paths normally used by information to flow throughout the Internet. Although traffic congestions and local network's policies may cause unpredictable path's inflation, the traceroute paths are optimized in order to be the shortest ones between the source and the target nodes. Therefore, standard theoretical models of traceroute's explorations assume that the probes follow one of the possible shortest paths between the source and the destination. More precisely, we can include in the path only one shortest path among all equivalent ones (either a priori fixed or randomly chosen), or all of them. All strategies can occur in realistic processes, that is probably a mixture of them, but people usually give special attention to the one with a unique choice of the shortest path between nodes, that clearly brings to the worst overall sampling. We will also consider this case.

A one-to-all traceroute process is thus represented by an iterative algorithm running on a given network, that starting from a single source generates a spanning tree to all other nodes. Multi-source processes consist in merging different single-source spanning trees. In general, the reliability of traceroute-like sampling methods strongly depends on both the number of sources deployed on the network and the level of degree heterogeneity [10]. In fact, Lakhina et al. [30] first showed numerically that sampling from single sources introduces uncontrolled biases and the observed statistical properties may sharply differ from the original ones. More recently, Clauset and coworkers [8,12] have pointed out that, because of the particular search procedure, a one-to-all traceroute tree has a power-law degree distribution $\tilde{P}_1(k) \propto k^{-\alpha}$ even if the underlying network is not scale-free. Actually, this was analytically proved only for homogeneous random graphs with fixed or Poisson degree distribution. In this case, the traceroute tree has a power-law distribution with exponent $\alpha = 1$ up to a cut-off equal to the average degree z. For networks with power-law degree distributions $P(k) \propto k^{-\gamma}$, they suggested that the observed one should still be power-law but with a different exponent $\alpha < \gamma$. Within the same framework, but in partial contrast with this thesis, Cohen et al. [13] have rigorously showed that in case of power-law networks, the bias on the exponent γ is negligible. Other recent studies, based on mean-field approaches corroborated by numerical simulations, confirm the overall reliability of these mapping techniques on scale-free graphs [10,31]. Nevertheless, the debate on the traceroute process is still open, the main issues concerning the relevance of the biases in single-source processes and the improvements obtained using multiple sources. In the following, we try to address both these subjects using the theoretical approach developed in the previous section.

The analytical results mentioned above are actually based on approximated models that partially overlook the dynamical character of the process. The method used in reference [8] to study one-to-all traceroutes is based on differential equations, but node sampling is essentially static. It assumes that a node of a given degree can appear with the same probability at any temporal step of the process. This is approximately true on homogeneous networks since the term $\frac{i_k(t)}{i(t)}$ in equation (4) is $\simeq 1$, but it cannot be extended to the case of heterogeneous networks. The formal approach introduced in reference [12] is more general, but it is still based on a uniformly random process, the so-called "exposure on the fly" technique. It implicitly assumes a fitness-like variable homogeneously distributed on the "stubs" of a network, that plays the role of the time at which a node is explored. We improve this approximation using the dynamical method exposed in Section 2 for both single-source and multi-source processes. Obviously the framework becomes more complex and calculations can be performed analytically only in some special cases.

3.1 Homogeneous networks

Let us consider the one-to-all traceroute exploration of an homogeneous random graph with Poisson degree distribution $P(k) = e^{-z}z^k/k!$. The original traceroute dynamics is discrete in time: at each temporal step, an interfacial node ℓ is randomly selected to spread out probes towards all its unknown neighbors; then all reached nodes are included in the interface, whereas node ℓ is moved to the bulk. A continuous-time approximation is defined grouping together N discrete steps and passing to differential equations in the limit of large size N.

As the degree distribution is peaked around the average value z, one can safely do the further approximation that all nodes behave identically, that is their temporal evolution is described by the mean-field densities $u(t)$, $i(t)$, $b(t)$ [8]. A selected interfacial node is connected to a unreached node with probability $p = z/N$, therefore in a temporal step, $p\, u(t)$ unreached nodes are visited and one node passes from the interface to the bulk. This process translates in the following system of equations [8]

$$\begin{cases} \frac{d}{dt}u(t) = -z\, u(t), \\[2mm] \frac{d}{dt}i(t) = +z\, u(t) - 1, \\[2mm] \frac{d}{dt}b(t) = +1. \end{cases} \tag{10}$$

The solution gives $u(t) = \exp(-zt)$, $i(t) = 1 - t - \exp(-zt)$ and $b(t) = 1 - t$, with t going from zero to a maximum value T, that is the first root of $i(t) = 0$. Moreover, since in the traceroute sampling the probes emerging from the interfacial nodes visit all unreached neighbors, the mean-field probability to reach a node is just $\rho[u(t)] \simeq u(t)$.

The network's homogeneity implies that the internal degree profile of interfacial nodes is approximately the same of the underlying network at all times. According

to this approximation, for a traceroute spanning tree on an homogeneous Poisson graph equation (4) reduces to

$$\tilde{P}_1(k+1) \simeq \sum_{\ell=k}^{\infty} \int_0^T \frac{dt}{T} P(\ell+1) \binom{\ell}{k} \left[e^{-zt}\right]^k \left[1 - e^{-zt}\right]^{\ell-k},$$
(11)

that is exactly the result obtained by Clauset and Moore [8].

The integral in equation (11) can be carried out noting that $T \simeq 1$ (for $z \gg 1$) and passing to the variable $u = u(t) = e^{-zt}$, with $du = -zudt$. Then, we can also easily perform the sum over ℓ,

$$\begin{aligned} \tilde{P}_1(k+1) &\approx \sum_{\ell=k}^{\infty} P(\ell+1) \binom{\ell}{k} \left[\frac{u^k}{zk} F_{2,1}(k, k-\ell, k+1, u)\right]_{e^{-z}}^1 \\ &\approx \frac{e^{-z}}{zk} \sum_{\ell=k}^{\infty} \frac{z^{l+1}}{l+1!} \binom{\ell}{k} \\ &\quad \times \left[\binom{\ell}{k}^{-1} - e^{-kz} F_{2,1}(k, k-\ell, k+1, e^{-z})\right] \\ &\approx \frac{1}{zk} \left[1 - \frac{z^k}{k!} e^{-z}\right] \simeq \frac{1}{zk}, \end{aligned}$$
(12)

where $F_{2,1}$ is the hypergeometric function, and the term proportional to e^{-kz} is negligible for sufficiently large values of k. As in reference [8], we get an observed power-law degree distribution $\tilde{P}_1(k) \sim k^{-1}$ [8], with a cut-off at $k \simeq z$. (It is worth to remark that Poisson random networks with average degree large enough to generate a power-law under sampling are extremely unlike in realistic systems, i.e. the Internet, community networks, etc. [10]).

The results obtained simulating the traceroute model from a single source on a Poisson random graph (e.g. using the unique shortest path algorithm proposed in Ref. [10]), confirm the k^{-1} behavior of the observed degree distribution (see Fig. 2A).

In Figure 2A, we also report the observed degree distributions obtained sampling from two or more sources, that clearly depart from a purely power-law shape. The observed weird behavior, neither power-law nor poissonian, in which a peak appears at low degree values and moves forward for increasing number of sources, can be easily understood, at a qualitative level, using the theory presented in Section 2. These peaks are due just to the superposition of power-law behaviors. In single-source experiments, most visited nodes have observed degree one, as they are discovered at the end of the process. The majority of them is rediscovered again in the same fashion during the second one-to-all process. Thus merging two spanning trees the overall distribution presents a peak at degree 2 instead of 1. The same happens for three sources, with a peak at degree 3, and for increasing number of sources. In general, the position of the peak k_p is not strictly equal to the number of sources ν, but $k_p \leq \nu$ usually holds. Solving numerically the recursive equation (9), we get the results reported in Figure 2B. The curves have the same behavior as in the simulations, characterized by a peak
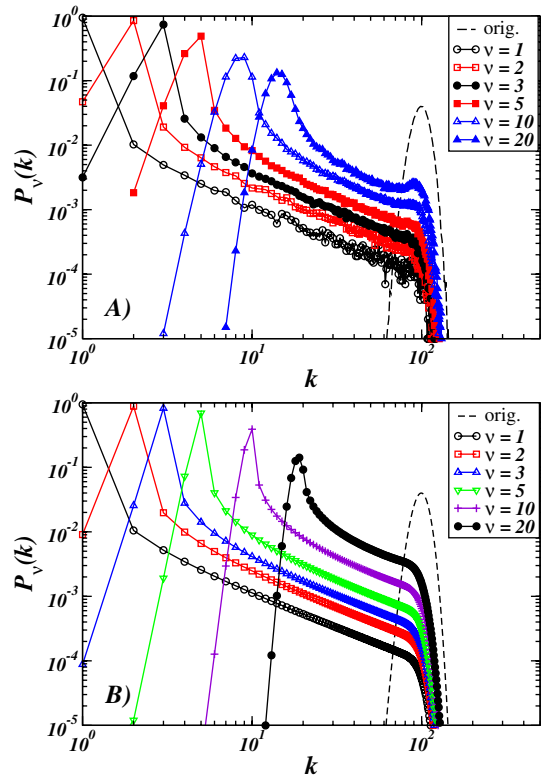
Fig. 2. A) Degree distribution $\tilde{P}_\nu(k)$ of the network produced by merging together ν spanning trees generated by one-to-all traceroutes algorithms on a Poissonian random graph of size $N = 10^5$ and average degree $z = 100$. Increasing the number ν of sources, i.e. of trees, the degree distribution becomes closer to the original one. A small peak around z is developed because of metric correlations. B) Degree distribution $\tilde{P}_\nu(k)$ obtained numerically from the recursion relation in equation (9) for a Poissonian random graph of average degree $z = 100$.

at increasing degree values, then a decrease up to a cut-off about $k \approx z$. Strikingly, the assumption of complete uncorrelation of successive spanning trees seems to be approximately correct for homogeneous random graphs. This is true up to a certain amount of sources (about 20 in Fig. 2B), above which the "metric" correlations between shortest paths are not negligible. At this point, a peak at the original average degree z is developed.

Our theoretical approach shows that the observed power-laws in one-to-all traceroutes on homogeneous random graphs are the result of a kind of convolution over a family of peaked symmetric distributions. This convolution process can be naturally "inverted" by increasing the number of observation points. For a large number of sources the tree merging process corresponds to another kind of convolution on the power-laws emerging from single-source experiments and produces an unbiased sampling of the original degree distribution. However, the minimal number ν^* of sources required to obtain an unbiased degree distribution is considerably large in homogeneous networks: from simple arguments and numerical evidences, $\nu^* \sim \mathcal{O}(z)$.
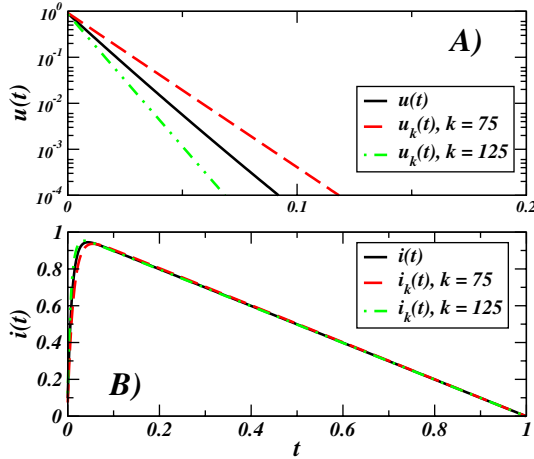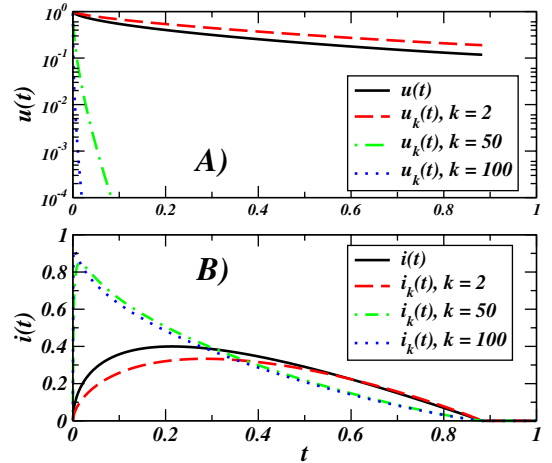
Fig. 3. Temporal behavior of the global and partial densities of unreached (A) and interfacial (B) nodes in an homogeneous random graph with Poisson degree distribution of average degree $z = 100$. The curves are obtained by numerical integration of the system in equation (13).

Fig. 4. Temporal behavior of the global and partial densities of unreached (A) and interfacial (B) nodes in a power-law random graph with exponent $\gamma = 2.5$ and average degree $z \simeq 4.5$. The curves are obtained by numerical integration of the system in equation (13).

In general the mean-field approximation on networks can be improved considering degree-dependent mean-field quantities, therefore we consider the system of differential equations describing the dynamics for degree-dependent partial densities of bulk, interfacial and unreached nodes. In uncorrelated networks, it reads

$$\begin{cases} \frac{d}{dt} u_k(t) = -\sum_h \frac{(h-1)}{z} P(h) \frac{i_h(t)}{i(t)} k \, u_k(t), \\[2mm] \frac{d}{dt} i_k(t) = +\sum_h \frac{(h-1)}{z} P(h) \frac{i_h(t)}{i(t)} k \, u_k(t) - \frac{i_k(t)}{i(t)}, \\[2mm] \frac{d}{dt} b_k(t) = +\frac{i_k(t)}{i(t)}, \end{cases} \quad (13)$$

where $\frac{(h-1)}{z} P(h) \frac{i_h(t)}{i(t)} k \, u_k(t)$ is the probability that emerging from an interfacial node of degree h we reach a node of degree k still unreached at time t. Note that the above system reduces to equations (10) when the network is a regular random network of degree distribution $P(k) = \delta_{k,z}$. For poissonian networks, the exact time-depending behavior of the partial densities is quite complicated, as evidenced by the curves reported in Figure 3 obtained solving numerically the system in equation (13). In the numerical solution we take initial conditions $i_h(0) = C_0 \delta_{h,z}$ with $C_0 \simeq \Delta t$, since both the initial condition and the temporal step Δt should be of order $1/N$ in a system of size N.

The general behavior can be explained with simple arguments. At the beginning of the process, the probability of having a node of degree k at the interface is purely topological, i.e. $\frac{i_k(0^+)}{i(0^+)} \simeq \frac{k}{z}$. The late times behavior for $t \gg 0$, can instead be computed knowing that the corresponding behavior of $i(t)$ is approximately linear, i.e. $i(t) \simeq 1 - t$, and that $u_k(t)$ decreases exponentially fast in time. These results, obtained plugging the short times approximation $\frac{i_k(t)}{i(t)} \simeq \frac{k}{z}$ into the equation for $\frac{du_k(t)}{dt}$, are verified in the numerics. Hence, from equation (13), $\frac{di_k(t)}{dt} \approx ke^{-kt} - \frac{i_k(t)}{1-t}$.

For sufficiently large k, the first term at the r.h.s. can be neglected, thus after integration we get $i_k(t) \simeq 1 - t$. Therefore, the quantity $\frac{i_k(t)}{i(t)}$ is expected to approach the unity for sufficiently large $t \gg t_k^*$, where t_k^* is the time at which the maximum value of $i_k(t)$ is reached. From the previous arguments one expects $t_k^* \propto 1/k$.

These simple calculations, and the numerical results reported in Figure 3, show that the short time behavior of the sampling process is not trivial at all. Nonetheless, the degree distributions obtained solving numerically the system in equation (13) and plugging the corresponding partial densities in equations (4) and (9) are in perfect agreement with the analytical results based on the approximation of complete homogeneity (not shown).

3.2 Heterogeneous networks

The degree inhomogeneity is instead expected to play an important role in the exploration of networks with skew and fat-tailed degree distributions, in which the dynamical sampling of nodes is far from being a uniform process. Though to obtain the exact form of partial and global densities it is necessary to solve the evolution equations explicitly, that is in general very difficult, the qualitative behavior of these quantities for large degree values ($\ell \gg z$) can be deduced with some approximate argument. We will show that the tail of the degree distribution of a power-law random graph is sampled with negligible bias, even in single-source experiments. This is mainly due to the fact that high degree nodes arrive at the interface of the process almost immediately, and their neighbors are fairly sampled with a probability that depends only weakly on the dynamics.

At the beginning of the process, nodes with large degree are preferentially sampled, i.e. $i_\ell(t)/i(t) \simeq \ell/z$, implying that the number of unreached hubs rapidly decays

to zero. At this point, almost all high degree nodes are at the interface of the process, $i_\ell(t)$ varies slowly, and the sampling among high degree nodes inside the interface becomes almost uniformly random (i.e. hubs of degree ℓ are picked with probability $P(\ell)$). Hence, for sufficiently large times and degrees $\ell \gg z$, the quantity $\frac{i_\ell(t)}{i(t)}$ is expected to become independent of the degree ℓ. Solving numerically equations (13), we find that, for $t \gg 0$, $\frac{i_\ell(t)}{i(t)} \propto t^{-\beta}$, with a cut-off close to the final sampling time T and the exponent β depending on the details of the degree distribution (not only on γ). The numerical results presented in Figure 5 suggest a scaling function for $\frac{i_\ell(t)}{i(t)}$ of the type

$$\frac{i_\ell(t)}{i(t)} \approx \frac{\ell}{z} \mathcal{F}\left[t\left(\frac{\ell}{z}\right)^{1/\beta}\right], \tag{14}$$

with $\mathcal{F}(x) \propto x^{-\beta}$ when $x \gg 1$ and $\mathcal{F}(x) \approx 1$ when $x \ll 1$. The scaling form is correct up to a time $T_\ell \sim T(\ell/z)^{-1/\beta}$, at which the quantity vanishes. Note that for a system of size N, and power-law degree distribution of exponent γ, the maximum degree scales as $\ell_{max} \sim N^{1/(\gamma-1)}$. Since the temporal step of the dynamics (integration step) is $\Delta t \sim 1/N$, a realistic minimum observation time for the above curves is $\Delta t^{1/(1-\gamma)}$, that is traced in Figure 5A with a dotted vertical line. It is clear that for high-degree nodes the power-law scaling behavior dominates the most of the dynamics. The result is surprising since one would expect that, in a power-law network, $i_\ell(t)/i(t) \simeq \ell/z$ during the whole dynamics. We will see that this behavior is crucial to get an unbiased estimate of the exponent.

The other interesting time-dependent quantity is the probability to visit an unreached node, $\bar{u}(t) = \sum_\ell \frac{\ell}{z} P(\ell) u_\ell(t)$. At long times and high degrees ℓ, the quantity $u_\ell(t)$ is non zero only for low degree nodes, so the temporal behavior of $\bar{u}(t)$ is similar to that of $u(t)$. Actually, it seems to decay from 1 with a law that is clearly slower than an exponential one, but faster than a power-law. Indeed, if $i_\ell(t)/i(t)$ follows a power-law behavior, formally integrating equations (13) one gets $\bar{u}(t) \sim e^{-at^\alpha}$, with $\alpha < 1$. However, numerical integration does not clarify the possible relation existing between α and β.

Plugging the above results into the sampling integral (Eq. (4)), it is actually possible to show numerically that the observed degree distribution maintains the same functional form of the original one, at least for high degree nodes. The comparison between observed degree distribution obtained from simulations and by solving numerically the sampling equations are reported in Figure 6. The qualitative behavior is the same: the traceroute sampling on power-law random graphs reproduces the original degree distribution without any significative bias.

Some further insights on the reason of this result can be obtained with the following rough argument. Let us
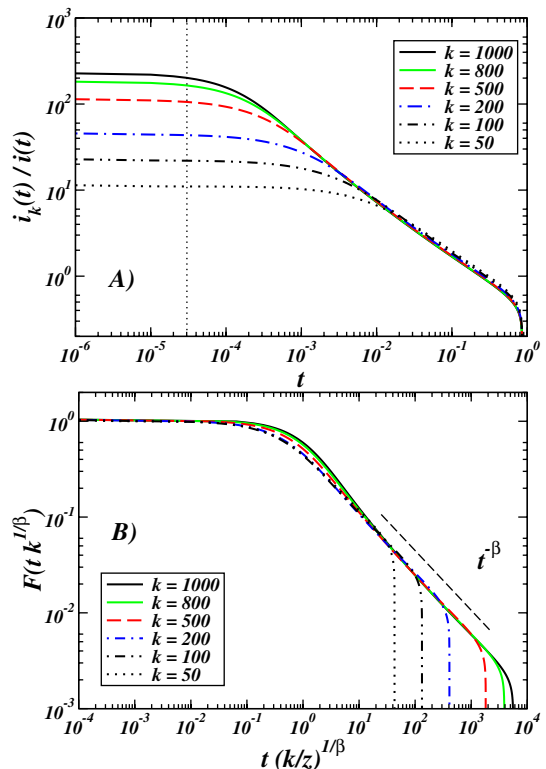
Fig. 5. A) Temporal behavior of the function $i_k(t)/i(t)$ for high-degree nodes in a power-law random graph. B) The scaling function $\mathcal{F}$ (see Eq. (14)) associated with some universal behavior of $i_k(t)/i(t)$ for high-degree nodes.

consider the sampling formula

$$\tilde{P}_1(k+1) \simeq \sum_{\ell=k}^{\infty} \int_0^T \frac{dt}{T} \frac{i_{\ell+1}(t)P(\ell+1)}{i(t)}$$
$$\times \binom{\ell}{k} [\bar{u}(t)]^k [1-\bar{u}(t)]^{\ell-k} \tag{15}$$

and approximate it for $\ell \gg z$ considering that 1) during most of the dynamics the quantity $i_k(t)/i(t)$ assumes a scaling form $\propto t^{-\beta}$, and 2) the binomial probability can be approximated by a Gaussian peaked around its maximum, i.e. $\bar{u}^* \sim k/\ell$. A rough estimate can be done performing the Gaussian integral at the saddle-point or, in an equivalent way, recalling a property of Dirac delta functions, i.e. $\int f(t)\delta(g(t))dt \approx \sum_i f(t_i)/|g'(t_i)|$, where t_i are the zeros of $g(t)$ and $g'(t_i)$ is the derivative of the function g in these points. For large k and ℓ, $g(t) \approx \ell\bar{u}(t) - k$, thus $|g'(t)| \propto \ell\frac{d\bar{u}}{dt}|_{t_\ell}$ (in which t_ℓ is the time at which $\bar{u}(t_\ell) \simeq u^* \simeq k/\ell$ for given values of k and ℓ). Putting all ingredients together, the sampling formula becomes

$$\tilde{P}_1(k+1) \simeq \sum_{\ell=k}^{\infty} P(\ell+1)\frac{1}{\ell+1}\frac{t_\ell^{-\beta}}{|\frac{d\bar{u}}{dt}|_{t_\ell}}. \tag{16}$$
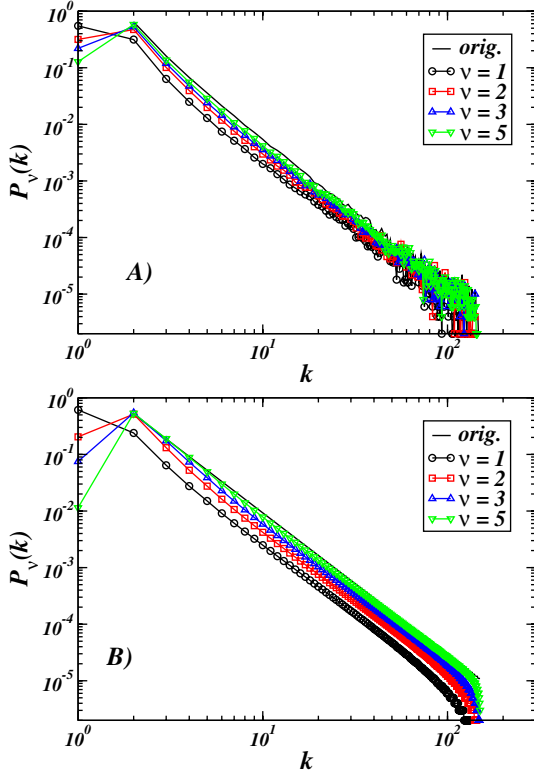
Fig. 6. A) Degree distribution $\tilde{P}_\nu(k)$ of the network produced by merging together ν spanning trees generated by one-to-all traceroutes algorithms on a power-law random graph of size $N = 10^4$, average degree $z \simeq 4.5$ and exponent $\gamma \simeq 2.5$. The level of sampling, almost unbiased with just a single source, improves increasing the number of sources. B) Degree distribution $\tilde{P}_\nu(k)$ obtained numerically from the recursion relation in equation (9) for a power-law random graph with average degree $z \simeq 4.5$ and exponent $\gamma \simeq 2.5$.

Since $\bar{u}(t) \sim e^{-at^\alpha}$, the derivative is proportional to k/ℓ times some power-law in time, and we get

$$\tilde{P}_1(k+1) \propto \sum_{\ell=k}^{\infty} P(\ell+1)\frac{1}{k}\left[\log(\frac{\ell}{k})\right]^{\frac{1-\beta-\alpha}{\alpha}}. \qquad (17)$$

For large degree we can neglect the logarithmic contributions in the sum over ℓ, finding $\tilde{P}_1(k) \propto k^{-\gamma}$.

According to this result, the degree distribution of the spanning tree emerging from one-to-all traceroute samplings of a scale-free graph (with exponent between 2 and 3) is qualitatively the same of that of the underlying network, at least for high degree nodes. The result is in agreement with a recent analysis by Cohen and coworkers [13], in which some rigorous bounds for the traceroute biases in power-law networks are obtained using the exposure technique. We believe that the general picture describing the behavior of traceroute-like processes could be extracted from the present approach in a much easier way than from the "exposure on the fly" method. Moreover, the framework can be straightforwardly extended to study, at least numerically, the effects of other relevant parameters, and correlations. In these perspective our ap-

proach can be considered as a benchmark that could be useful in the problem of network inference [33] and bias reduction [34].

4 Conclusions

The study of the interplay between topological and dynamical properties of networks is of primary interest in the current research on complex systems. With the present work, we introduce in this framework a general method to investigate the topological properties of growing clusters that are dynamically defined by a given class of spreading processes and algorithms. These are processes that starting from a single source span the whole network, traversing all nodes only once. The temporal evolution is defined by means of a system of differential equations for the (partial) densities of bulk, interfacial, and unreached nodes.

Our approach, that allows to compute analytically or numerically the degree distribution of the emerging tree-like structure, is based on the idea that during its evolution the dynamics performs a sampling of the local structure of the underlying network. As the sampling rate depends on the dynamical properties themselves, the degree distribution of the emerging subnetwork may differ considerably from that of the original network. The generalization to study multiple-source processes is also discussed, at least in the approximation that the overall structure is obtained merging together collections of independent single-source processes.

A natural application of the method presented here consists in the analysis of the dynamical sampling of networks. We have provided a deeper insight in the qualitative behavior of traceroute-like processes, shedding light on the dynamical mechanism at the origin of the observed topology. The reasons for the different sampling accuracy observed in homogeneous and heterogeneous networks should be now clear. In particular, we have shown that on homogeneous networks,

- in a single source sampling, the network is observed to have a power-law degree distribution with exponent -1 up to a cut-off about the original average degree z;
- increasing the number of sources destroys the power-law behavior, but a fair sampling requires about $\mathcal{O}(z)$ sources;
- "metric" correlations (that can be associated to the betweenness centrality of nodes) favor a better sampling of the local topology for increasing number of sources.

In contrast, on heterogeneous networks,

- high-degree nodes are preferentially sampled at the beginning of the process, and with higher accuracy (in this case high-degree nodes are essentially those with highest betweenness);
- in single-source experiments, power-law degree distributions are sampled with negligible biases;
- the overall sampling of the degree distribution becomes very accurate with just few sources.

We believe that a good understanding of these kind of processes is fundamental in order to improve the performances of current dynamical sampling techniques applied to technological networks such as the Internet and the Web.

As mentioned in the Introduction, there are other dynamical processes that can be described using the present approach, from the epidemic spreading, to broadcast trees and search techniques. We just mention a couple of cases that may find further developments. One consists in a recently proposed model of search in social networks, in which the search efficiency decays with the distance [35]. Let us consider an uncorrelated homogeneous random network, on which we perform a distance dependent snowball search such that the probability to visit a still unreached neighbor of an interfacial node is $\theta(t) \propto (t + A)^{-\beta}$, with $\beta \in [0, 1)$. The corresponding global density is obtained solving the equation $\frac{du(t)}{dt} = -z\theta(t)u(t)$. The limit in which $\beta \to 1$ is particularly instructive, since the global density decreases much slower than an exponential. Performing the calculations, $u(t) \sim A/(A+t)$ and the emerging tree presents a degree distribution $\tilde{P}_1(k) \approx k^{-2}$ up to a cut-off around z. The example shows how easy can be to find processes that generate power-law degree distributions out of exponential networks.

The second relevant example concerns epidemic-like spreading phenomena. Here the system of differential equations governing the dynamics is the celebrated susceptible-infected-removed (SIR) model [36], in which unreached nodes are identified with susceptible nodes, interfacial nodes with infected ones and the bulk nodes correspond to removed individuals. In analogy with the traceroute dynamics, we can write a system of first-order differential equations governing the temporal evolution of partial densities [37]. The fundamental difference between the two dynamics is that in the traceroute model at each time step the algorithm chooses an interfacial node to sample its neighbors, while in the SIR model all infected nodes have a fixed probability to spread the virus to their neighbors (with spreading rate λ). However, for small values of λ (above the percolation threshold) the topological structure is still tree-like and can be analyzed as in Section 2. When different transmission properties (e.g. degree-dependent and distance-dependent rates λ) are taken into account, the growing infection region might display very non-trivial topologies. At the same time, the knowledge of the topological structure of an infection's outbreak may be relevant for the design of more appropriate immunization strategies.

From a general point of view, the present formalism can be extended to study (at least numerically) the effect of degree-degree correlations or quenched disorder, that have not been considered here but play an important role in all real experiments.

In conclusion, we expect that analyses like the one performed in this paper will allow to get a better understanding of the functional interplay between a network and the dynamical processes evolving on it.

References

1. *Handbook of Graphs and Networks: From the Genome to the Internet*, edited by S. Bornholdt, H.G. Schuster (Wiley-VCH, Berlin, 2002)
2. S.N. Dorogovtsev, J.F.F. Mendes, *Evolution of networks: From biological nets to the Internet and WWW* (Oxford University Press, Oxford, 2003)
3. R. Albert, A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002)
4. M.E.J. Newman, SIAM Rev. **45**, 167 (2003)
5. S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, Phys. Rep. **424**, 175 (2006)
6. T. Petermann, P. De Los Rios, Eur. Phys. J. B **38**, 201 (2004)
7. M.P.H. Stumpf, C. Wiuf, R.M. May, PNAS **102**, 4221 (2005); M.P.H. Stumpf, C. Wiuf, Phys. Rev. E **72**, 036118 (2005)
8. A. Clauset, C. Moore, Phys. Rev. Lett. **94**, 018701 (2005)
9. S.H. Lee, P.-J. Kim, H. Jeong, Phys. Rev. E **73**, 016102 (2006)
10. L. Dall'Asta, J.I. Alvarez-Hamelin, A. Barrat, A. Vázquez, A. Vespignani, Phys. Rev. E **71**, 036135 (2005); L. Dall'Asta, J.I. Alvarez-Hamelin, A. Barrat, A. Vázquez, A. Vespignani, Lect. Notes in Comp. Sci. **3405**, 140 (2005); L. Dall'Asta, J.I. Alvarez-Hamelin, A. Barrat, A. Vázquez, A. Vespignani, Theor. Comp. Sci. **355**, 6 (2006)
11. R. Pastor-Satorras, A. Vespignani, *Evolution and structure on the Internet: a statistical physics approach* (Cambridge University Press, Cambridge, 2004)
12. D. Achlioptas, A. Clauset, D. Kempe, C. Moore, in STOC'**05**, 694 (2005)
13. R. Cohen, M. Gonen, A. Wool, preprint arXiv:cs.NI/0611157 (2006)
14. R. Castro, M. Coates, G. Liang, R. Nowak, B. Yu, Statistical Science **19**, 499 (2004)
15. T. Kalisky, R. Cohen, O. Mokryn, D. Dolev, Y. Shavitt, S. Havlin, Phys. Rev. E **74**, 066108 (2006)
16. G. Pant, P. Srinivasan, F. Menczer, *Web Dynamics* edited by M. Levene and A. Poulovassilis (Springer, 2004)
17. R.M. Anderson, R.M. May, *Infectious Diseases of Humans: Dynamics and Control* (Oxford Univ. Press, Oxford, 1991)
18. Y. Moreno, M. Nekovee, A.F. Pacheco, Phys. Rev. E **69**, 066130 (2004); M. Nekovee, Y. Moreno, G. Bianconi, M. Marsili, Physica A **374**, 457 (2007)
19. E. Rogers, The diffusion of innovations (Free Press, New York, 1995). T.W. Valente, *Network Models of the Diffusion of Innovations* (Hampton Press, Cresskill, NJ, 1995)
20. M. Granovetter, Amer. J. Soc. **83**, 1420 (1978)
21. M.E.J. Newman, Phys. Rev. E **66**, 016128 (2002)
22. M. Boguñá, R. Pastor-Satorras, Phys. Rev. E **66**, 047104 (2002)
23. N.C. Wormald, *Ann. Appl. Probab.* **5**, 1217 (1995); N.C. Wormald, in *Lectures on Approximation and Randomized Algorithms*, edited by M. Karonski, H.J. Proemel, (Warsaw, 1999), pp. 73–155
24. M. Molloy, B. Reed, Comb., Prob. and Comp. **7**, 295 (1998)

25. D. Achlioptas, Y. Peres, J. Amer. Math. Soc. **17**, 947 (2004)
26. See e.g. the recent preprint, R. Oliveira, B. Zhang, and L. Zhang, "Observing the Evolution of Internet AS Topology", to appear in ACM SIGCOMM, Kyoto, Japan (2007)
27. J. Pansiot, D. Grad, ACM Computer Communication Review **28**, 41 (1998)
28. R. Govindan, H. Tangmunarunkit, IEEE INFOCOM **2000**, 1371 (2000)
29. M. Faloutsos, P. Faloutsos, C. Faloutsos, Comput. Commun. Rev. **29**, 251 (1999)
30. A. Lakhina, J. Byers, M. Crovella, P. Xie, Proc. IEEE INFOCOM 2003 (2003)
31. J.-L. Guillaume, M. Latapy, Complex Systems **16**, 83 (2005); J.-L. Guillaume, M. Latapy, D. Magoni, Computer Networks **50**, 3197 (2006)
32. Y. Shavitt, E. Shir, cs.NI/0504107
33. F. Viger, A. Barrat, L. Dall'Asta, C. Zhang, E. Kolaczyk, Phys. Rev. E **75**, 056111 (2007)
34. A.D. Flaxman, J. Vera, preprint, `arXiv:cond-mat/0705.3243` (2007)
35. F. Wu, B.A. Huberman, L.A. Adamic, J.R. Tyler, Physica A **337**, 327 (2004)
36. W.O. Kermack, A.G. McKendrick, Proc. Roy. Soc. Lond. A **115**, 700 (1927)
37. Y. Moreno, R. Pastor-Satorras, A. Vespignani, Eur. Phys. J. B **26**, 521 (2002)