A Realistic Model for Complex Networks

Jean-Loup Guillaume, Matthieu Latapy
LIAFA, University of Paris 7
2 place Jussieu
75005 Paris
France.
(guillaume,latapy)@liafa.jussieu.fr

***Abstract*—It appeared recently that the classical random network model used to represent complex networks does not capture their main properties (clustering, degree distribution). Since then, various attempts have been made to provide network models having these properties. We propose here the first model which achieves the following challenges: it produces networks which have the three main wanted properties, it is based on some real-world observations, and it is sufficiently simple to make it possible to prove its main properties. We first give an overview of the field by presenting the main models introduced until now, then we discuss some remarks on some complex networks which lead us to the definition of our model. We then show that the model has the expected properties and that it can actually be seen as a general model for complex networks.**

***Index Terms*—topology, generator, simulation, networks, graphs, modeling**

I. Introduction.

In a random network [1], [2] with n nodes, each of the $\frac{n\cdot(n-1)}{2}$ possible links exists with a given probability p. In other words, a random network is constructed from n nodes by choosing $m = p \cdot \frac{n\cdot(n-1)}{2}$ links at random. Until recently, this model was merely the only one available for the study of complex networks. It has been widely used in many contexts as various as the simulation of virus propagation in a population [3], [4], the evaluation of the impact of failures and attacks in computer networks [5], [6], the evaluation of the performance of algorithms on networks [7], [8] or computer networks protocols [9], etc.

However, it has been shown recently that most complex networks have some specific properties not captured by this model. Let us consider in particular the following three properties:

- the **average distance** is the average, over all the couples of nodes, of the distance between them, *i.e.* the minimal number of links one has to cross to go from one node to the other.
- the **clustering coefficient** is the probability of existence of a link between two nodes when they are both neighbors of a same node. It is computed by dividing the total number of triangles (trios of nodes with all the three possible links) in the network by the total number of connected triples (trios of nodes with at least two links).
- the **degree distribution** is the function P_k giving the proportion of nodes with degree exactly k, *i.e.* with exactly k neighbors, in the network. In other words, P_k is the probability that a randomly chosen node has degree k.

In a random network as described above, it is known that the average distance grows as $\log(n)$ [2]. Moreover, the clustering coefficient is equal to p since each pair of nodes is connected with the same probability p. This means that, if one considers a family of networks where the average degree is a constant (which is reasonable in the real-world cases), then the clustering coefficient tends to 0 when n grows. Finally, the degree distribution follows a Poisson law, which implies in particular that the number of nodes with a degree k decays exponentially with the difference between k and the average degree.

In most complex networks [10], [11], [12], the average distance also grows as $\log(n)$. However, the clustering coefficient is several orders of magnitude larger than in random networks (it is in general considered as independent of n). Moreover, the degree distribution follows a power law $P_k \sim k^{-\beta}$ which implies that despite the fact that most nodes have a low degree, there exists few nodes with very large degree. In other words, the number of nodes with a degree k decays polynomially (not exponentially) with k.

These properties have been measured in a wide variety of complex networks, including biological networks (cellular networks, protein interactions, dependencies between species, etc.), social networks (acquaintance, citations of papers, phone calls, exchanges of e-mails, Web links, etc.), technical networks (interconnection of routers or AS on the Internet, energy delivery network, peer-to-

networks	nodes	links	C	C_{rand}
Internet	228263	320149	0.06	0.00001
Web	325729	1090108	0.466	0.00002
Actors	392340	15038083	0.785	0.0002
Co-authoring	16401	29552	0.638	0.0002
Co-occurrence	9297	392066	0.822	0.009
Protein	2113	2203	0.153	0.001

Fig. 1. The main statistics for the examples we will use in this paper. For each network, we give its number of nodes, its number of links, its clustering coefficient, and finally the clustering coefficient of a typical random network with the same number of nodes and links.

peer networks, etc.), and many others. All these networks display the same behavior concerning the three properties we cited, which makes them very different from random networks. See [10], [11], [12] for more information on the networks satisfying these properties.

These results are true in particular for the following set of complex networks, which we will take all along this paper as a representative set of examples for our experiments. We choose them because they span quite well the large variety of complex networks we have cited.

- **Internet topology.** It represents the interconnection of routers (or autonomous systems) on the Internet. We will use various explorations of this network from [13], [14].
- **Web graph.** It is composed of the Web pages and the hyperlinks between them. We will use here the Notre Dame Web Web graph from [15].
- **Actors graph.** Two actors are connected if they play together in a movie. This network is widely studied because it is easily available through the Internet Movie Database [16].
- **Co-occurrence graph.** Two words are connected if they appear in a same sentence of a given book. Here, we will use a version of the Bible [17].
- **Co-authoring graph.** Two persons are linked if they have signed a paper together. We will use such a network obtained from the Los Alamos preprint archive [18].
- **Protein graph.** In [19] the authors link together two proteins of a given biological system if they influence each other. We will consider this example too, using networks from [15].

The main properties of these complex networks are summarized in Figure 1. Notice that, as announced, they all have a very low average distance, a power law distribution of degrees, and a high clustering, which makes them significantly different from random networks with the same number of nodes and links.

In many contexts, the properties of the underlying topology have a strong influence on the phenomena of interest. It has for example been shown that the robustness of systems like the Internet to failures and attacks highly depends on characteristics of its topology like its degree distribution [5], [6]. Likewise, the spreading of viruses or rumors can be accelerated or stopped using properties of the underlying social topology [3], [4], [20]. It also has been shown that the properties we cited have an impact on the performance of the protocols and algorithms [7], [8], [9], and we are only at the beginning of the investigation of the consequences of complex networks properties in many contexts. Therefore, it is important for the relevance of the simulation results, as well as the theoretical ones, to use realistic topologies.

This is why various models have been proposed since it appeared that the classical random network model does not fit the main properties met in practice. We propose an overview of the main such models in Section II. However, as we will discuss in this section, these models either fail to capture one of the three properties cited above, or fail to give an intuitive and realistic interpretation of the origin of these properties. The core of this paper is the introduction of a new model which captures the three properties cited, based on some remarks on how some complex networks are *really* constructed. These remarks, as well as the model and some proofs of its properties, are presented in Section III. Finally, we show in Section IV that this model can be seen as very general, and we finish the paper in Section V by a discussion of directions for further investigations pointed out by our work.

II. Context

Many models of complex networks have already been proposed in the literature, mainly since it has been observed that the classical random model is not suitable for the modeling of complex networks. Some of them attempt to explain the general properties we have cited. Others are aimed at reproducing real world construction processes and generally model a specific network of special interest. We give in this section an overview of the field by presenting the most famous models for general complex networks in a first step, and then the main models specifically designed for the Internet topology. This provides a description of the context in which we propose our new model, and how it may be related to previous works.

A. Generic models

The first generic model of realistic complex networks, which is also the most famous one, has been introduced in 1998 by Watts and Strogatz [21]. As shown in Figure 2,

one starts with a ring of n nodes in which each node is connected to its k nearest neighbors, for a given k. Then, each link is rewired with probability p by choosing randomly a new extremity.

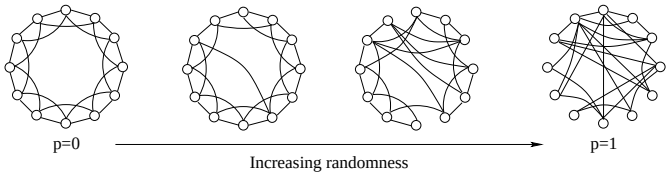

Fig. 2. The Watts and Strogatz model: from order (high clustering coefficient, high average average distance) to randomness (low clustering coefficient, low average distance). In between, the networks have both a high clustering coefficient and a low average distance, which captures a property of complex networks

Simulations of this model confirm the basic following intuition: the average distance is high (linear in n) if p is small, since only a few links are rewired and so the network is almost a ring. Notice however that, since each node is connected to its nearest neighbors, these neighbors are linked together and so the clustering coefficient is high. On the other hand, if p is high, then almost all the links are rewired, and so the network is similar to a random network: the average distance is low and so is the clustering coefficient. For medium values of p, the network has both a small average distance and a high clustering coefficient, which corresponds to the first two general properties of complex networks we have cited. Notice however that all these properties have been verified experimentally but no formal proof has been given, except for the exact value of the average distance [22].

Another important step was done when Albert and Barabási introduced their model based on *preferential attachment* [23], [24]. The idea can be well understood if we think about the way new Web pages connect to older ones. Intuitively, when you create a new Web page, you will more likely connect it to a well known one rather than a randomly chosen one. Since a page tends to be more famous when it has more links pointing to it, a new Web page tends to connect to well connected Web pages.

This "rich gets richer" or "popularity is attractive" principle can be derived in a model where nodes arrive one by one in a network and choose their neighbors with a probability function depending on the degree (a polynomial in the degree for instance). This simple model has been studied a lot and is now well know (refer to [10] for a survey of its properties). For instance, the degree distribution of the nodes follows a power law whose parameter can be controlled by the probability function. The average distance of such a network is logarithmic in the number of nodes, and the clustering coefficient is quite low (going to 0 with the number of nodes). This last point is annoying, but one has to recall that the preferential attachment is more a concept than a model by itself. This concept can therefore be used in other models as a simple and natural scheme in order to get a power law distribution for the degree. In particular we are going to use it in the model we will introduce in the next section.

Both Watts and Strogatz model and the Albert and Barabási one have been introduced to model generic behavior of complex networks. However, they both fail in producing networks having each of the three properties we cited. Others models have been introduced which achieve this goal [25], [26], but they are based on artificial processes which cannot be considered as realistic. Some specific networks, like the Internet, have also lead to specific models because of their prime importance. We give an overview of these models in the case of the Internet in the next section.

B. Internet specific models

Modeling the Internet topology is of prime interest for many purposes ranging from simulation to network management, or the development of specific algorithms (QoS routing, group communication, etc). In this context, attempts to model the specificities of the Internet topology go back to 1988 with the Waxman model [27] (we deliberately omit to cite random networks model as an Internet model). Hereafter we are going to present briefly some of the main models introduced in the last 15 years. This gives an idea of how the field has evolved during this time, and how our work may be inserted in this evolution. For surveys and discussions on these topics, we refer to [28], [29].

- Waxman model [27]: n nodes are placed in an Euclidean space, two of them being linked with a probability $\alpha \cdot e^{-d/\beta L}$, where d is the Euclidean distance between the nodes, L is the diameter and α and β are two parameters of the model: α regulates the number of links, while β regulates the ratio between the number of short links and long links.
- Hierarchical model from Zegura et al. [28]: each node of a network obtained by any given model (Waxman for instance) is expanded into a local network. This process can be iterated more than once.

In 1999, Faloutsos et al. [30] gave evidence of the fact that many power laws appear naturally in the description of the Internet topology. They show in particular that the degree distribution follows such a law, which was not expected before. This discovery made clear that previous models were not well suited to represent the reality of Internet topology, since they do not produce networks with

power law degree distributions. Therefore, some efforts have been made to give more realistic models, in particular to this respect.

- ACL [31] is a basic model which generates a random network with prescribed distribution of degrees. The model assigns to each node a degree drawn from the distribution. Then, each node is duplicated as many times as its degree, and finally pairs of nodes are chosen randomly to create the network. This simple scheme is very general and we will use it in Section III.
- BRITE [32] divides a square in a number of sub-squares (like a square grid), and assigns a number to each of them following any distribution (generally Poisson or power-law). This is the number of nodes in the sub-square. Then each node is placed randomly in each sub-square, and the links are added following a preferential connectivity and/or a preferential local connection which allows various behaviors. This model is aimed at modeling AS level Internet topology.
- INET [33] generates a network of a given size using some equations obtained from the measured evolution of the Internet AS level topology from 1997 to 2000. This generator incorporates rules similar to preferential attachment.
- GPL [29]: This model is similar to the one of Albert and Barabási above, but at each step either some links are added following a preferential attachment rule, or a new node is added and linked using preferential attachment. This model creates networks which have a higher clustering than the one of Albert and Barabási but it is still low and no formal proof has been given.
- HOT [34]: The first node of the network plays a special role and is called the source. Nodes are placed one by one randomly on the unit square and each new node is linked to the previously existing one which minimize a linear function of the Euclidean distance to this old node and its hop-distance to the source. This model generates a tree but the linear function can be chosen in order to get a power law degree distribution.

Notice also the original work [35] which is devoted to the sampling of a sub-network of a given Internet topology with properties similar to the ones of the original network.

Many other attempts have been made to reach the goal of obtaining models which give networks having each of the three properties we have cited. Most of them are described in [10]. However, all these models fail to give an intuitive realistic and simple interpretation of the causes of the observed properties.

III. The Model

A bipartite network is a triple $G = (T, B, E)$ where T and B respectively contain the Top and Bottom nodes of the network and $E \subseteq T \times B$ contains the links. The difference between classical (unipartite) networks lies in the fact that links are allowed only between top nodes and bottom nodes.

Some of the complex networks we have described in Section I display a natural bipartite structure. For example, let us consider the Actors graph (two actors are linked if they are part of a given cast). If we define T as the set of films and B as the set of actors, then one can view this complex network as a bipartite network where each actor is linked to the films he/she played in (and therefore each film is linked to the actors of its cast).

The Co-authoring graph can also be viewed this way with T being the set of articles and B the set of authors, each author being linked to the papers he/she co-signed. Likewise, in a Co-occurrence graph, one can link each sentence to the words it contains.

Given such a bipartite network $G = (T, B, E)$, one can easily construct its unipartite version as follows: $G' = (B, E')$ where $E' = \{\{u, v\} | \exists t \in T, \{t, u\}, \{t, v\} \in E\}$. From the bipartite versions of the Actors graph, Co-authoring and Co-occurrence graphs, one can then find back their original (unipartite) versions as defined in Section I. In this unipartite version of the network, each top node induces a clique (complete subgraph) between the bottom nodes to which it is linked. See Figure 3.

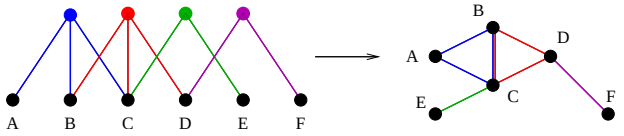

Fig. 3. A bipartite network and its unipartite version. Notice that the link $\{B, C\}$ is obtained twice since B and C have two neighbors in common in the bipartite network.

Since several complex networks we have cited in Section I have a natural underlying bipartite structure, one may wonder if their properties (like the clustering coefficient) are some consequences of this structure. This section is devoted to this idea. To explore it, we will deepen the study of the three bipartite complex networks we cited (Actors, Co-authoring and Co-occurrence). Then we will introduce our model, which is nothing but the random bipartite networks with prescribed degree distributions. We will study the properties of this model both analytically and experimentally, which will lead us to the conclusion

that the properties of the three complex networks we have cited can indeed be viewed as consequences of their natural underlying bipartite structure. The cases of networks which do not display such a structure is considered in Section IV.

A. *Real-world bipartite structure*

Two distributions can be naturally derived from a bipartite network: the top degree distribution and the bottom one. Figures 4 and 5 present these distributions for the Actor, Co-authoring and Co-occurrence networks.

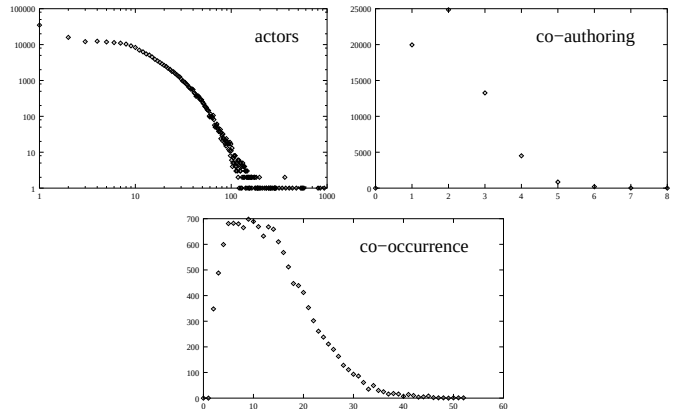

Fig. 4. Top degree distribution for the Actor, the Co-authoring and the Co-occurrence graphs

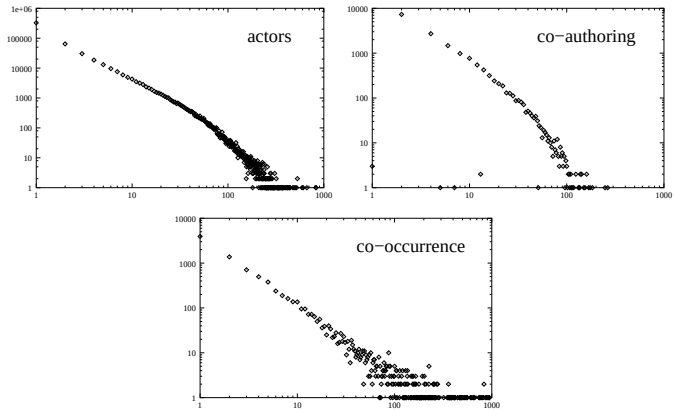

Fig. 5. Bottom degree distribution for the Actor, the Co-authoring and the Co-occurrence bipartite networks

At least for Co-occurrence and Co-authoring graphs, the top degree distribution is not heavy tailed, and exhibits a Poisson behavior. For the Actor graph, there is a significant number of top nodes with a high degree and the tail of the distribution looks like a power law. On the other hand, the top degree distributions iare heavy tailed for all the networks.

The degrees of a bottom node in the bipartite network (Figure 5) and the unipartite version (Figure 6) of the same network are not clearly related even if both distributions display a power law behavior. Actually the real degree of an bottom node is the sum of the degrees of the top nodes to which it is connected to, minus the overlap between the neighborhood of these nodes. Even if this notion of overlap is not formally defined, one can easily be convinced that it has a great impact on the degree distribution.

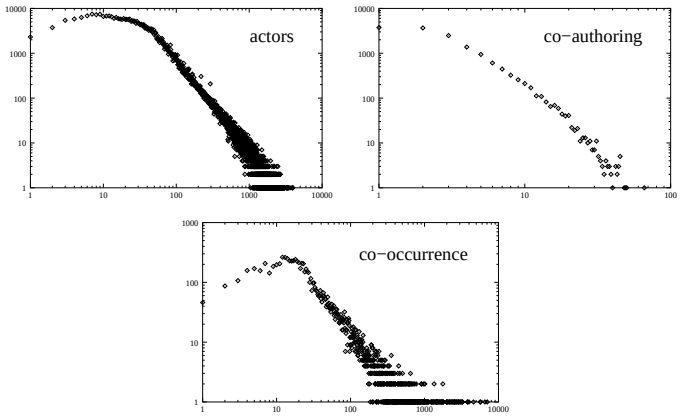

Fig. 6. Degree distribution in the unipartite version for the Actor, the Co-authoring and the Co-occurrence bipartite networks

At least for the two first networks, one can wonder why the degree of the unipartite network is power law distributed. Is it due to the degree distribution of the top nodes (clearly a power law), to the existence of top nodes with a high degree, or some more complex behaviors ? We will answer this questions below.

B. *The random bipartite model*

The model we propose actually is nothing but random uniform sampling of bipartite networks with prescribed top and bottom degree distribution. Such a network can be constructed as follows (see Figure 7):

1) generate both top and bottom nodes and assign to each node a degree drawn from the given distributions,
2) create for each node as many connection points as its degree,
3) link top and bottom connection points randomly,

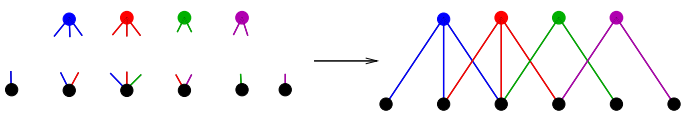

Fig. 7. Construction of a random bipartite network with prescribed degree distribution: first top and bottom nodes are drawn and each node is assigned a degree with respect to the given distributions, then links are chosen randomly between the two sets.

The bipartite model as presented before assumes that two distributions for both top and bottom nodes are explicitly given. We can also use some previous remarks to define these distributions implicitly. For instance we have noticed that the bottom degree distribution generally follows a power law, whereas top degree distribution is often a Poisson law. Therefore we could use a concept similar to preferential attachment to obtain a model in which we would first create the two sets of nodes and then at each step choose uniformly at random a top node, choose a bottom node according to its degree, and link both nodes. At each step the bipartite network has the required degree distributions (top Poisson law and bottom power law). Other processes may be created by merging some existing models with the bipartite one, but this paper is centered on the bipartite model with given degree distributions.

C. Properties

In this section, we give formal proofs for the main properties of the bipartite model. These results give a precise intuition on how and why the underlying bipartite structure implies the observed properties.

Let us denote by $N(v)$ the set of nodes adjacent to v, and by $N(u,v)$ the set of nodes adjacent to the link $\{u,v\}$, defined as $N(u) \cap N(v)$.

Degree distribution

Let us first consider the degree distribution of the unipartite version of a random bipartite network $G = (T, B, E)$. Given a bottom node u, we denote by $d(u)$ the degree of u in the bipartite network, and by $d_U(u)$ its degree in the unipartite network. We want to study the distribution of $d_U(u)$.

Lemma III.1: Let us consider a bottom node $u \in B$. The number of bottom nodes which have a neighbor (in T) in common with u, *i.e.* $d_U(u)$, is:

$$\frac{d(u)}{|T|} \cdot \sum_{t \neq u} d(t) + \mathcal{O}\left(\frac{d(u)^2}{|T|^2} \cdot \sum_{t \neq u} d(t)^2 \right)$$

Proof: The exact value of $d_U(u)$ is given by:

$$d_U(u) = \sum_{t \neq u} \left(1 - \frac{\binom{|T|-d(u)}{d(t)}}{\binom{|T|}{d(t)}} \right)$$

since the probability that a given bottom node t has a top neighbor in common with u depends only on the degree of both nodes and the number of top nodes. To simplify this formula, we can approximate the ratio $\binom{|T|-d(u)}{d(t)} / \binom{|T|}{d(t)}$ as follows:

$$\begin{aligned} \frac{\binom{|T|-d(u)}{d(t)}}{\binom{|T|}{d(t)}} &= \frac{(|T|-d(u))!(|T|-d(t))!}{|T|!(|T|-d(u)-d(t))!} \\ &\sim \frac{(|T|-d(t))^{d(u)}}{|T|^{d(t)}} \\ &\sim 1 - \frac{d(t)d(u)}{|T|} + \mathcal{O}\left(\left(\frac{d(t)d(u)}{|T|} \right)^2 \right) \end{aligned}$$

Therefore:

$$\begin{aligned} d_T(u) &\sim \sum_{t \neq u} \left(\frac{d(t)d(u)}{|T|} + \mathcal{O}\left(\left(\frac{d(t)d(u)}{|T|} \right)^2 \right) \right) \\ &\sim \frac{d(u)}{|T|} \sum_{t \neq u} d(t) + \mathcal{O}\left(\frac{d(u)^2}{|T|^2} \sum_{t \neq u} d(t)^2 \right) \end{aligned}$$

which gives the formula of the claim. ■

This lemma makes it possible to compute the probability for a node t in the unipartite network to have a given degree k if the bottom degree distribution is a power law with exponent β:

$$\begin{aligned} P[d_U(u) = k] &\sim P[d(u) = \frac{n}{\sum_{t \neq u} d(t)} \cdot k] \\ &\sim \frac{1}{(\sum_{t \neq u} d(t)) \cdot k)^\beta} \sim k^{-\beta} \end{aligned}$$

Therefore, as long as bottom degree distribution follows a power law, the degree distribution in the unipartite version of the network also follows a power law with the same exponent, which is indeed the case in practice as one can check in Figures 12 and 13.

Average distance

To study the average distance in the unipartite version of a network obtained with the model, we will use a result from L. Lu about the diameter (*i.e.* the largest distance between any two nodes) of some specific random networks:

Theorem III.2—[36]: Let $G = (V, E)$ be a network whose nodes are weighted with weights $w_1, \cdots, w_n$, such that each link $\{i, j\}$ appears with probability $w_i \cdot w_j \cdot p$. If the degrees of the nodes in V follow a power law with an exponent β strictly greater than 2, then the diameter of the network G is almost surely $\Theta(\log(n))$.

This theorem, together with the one presented above on the degree distribution of the unipartite version of the network, gives a way to prove that the diameter of the unipartite network scales with the logarithm of the size of the network.

Theorem III.3: Let $G = (T, B, E)$ be a bipartite network such that the bottom degree distribution follows a

power law with an exponent greater than 2, then the average distance of the unipartite version of G is almost surely $\mathcal{O}(\log(|B|))$.

Proof: Given two bottom nodes u and v in B, the probability that they are connected in the unipartite version is equal to the probability that they are both linked to a same top node in G. This probability is exactly proportional to $d_B(u) \cdot d_B(v)$. Therefore we can apply Theorem III.2 considering that the weight of each node is its degree and so the connection probability is ensured.

The unipartite version of the network therefore has all the properties necessary to apply Theorem III.2 as long as bottom degree distribution follows a power law with an exponent β strictly greater than 2. The diameter of the unipartite version of the network is almost surely $\mathcal{O}(\log(|B|))$, and since the diameter is an upper bound of the distances for each couple of nodes, the average distance also scales logarithmcally. ∎

Clustering coefficient

We will now give a lower bound for the clustering coefficient of a network obtained using the bipartite model. Recall that the clustering coefficient of a node v in a network $G = (V, E)$ is the probability that two of its neighbors are linked [21]:

$$cc(v) = \frac{|\{\{x, y\} \in E, x, y \in N(v)\}|}{\binom{d(v)}{2}}.$$

This coefficient is averaged over all the nodes to get the clustering coefficient of a network, which is equivalent [10] to the definition given in Section I. We are going to give a bound for the clustering coefficient of a node $b \in B$, denoted by $cc(b)$, in the unipartite version $G' = (B, E')$ of a bipartite network $G = (T, B, E)$ obtained with the model.

Two steps are used to achieve this. First we notice that the clustering coefficient of b depends only on the number of top nodes it is connected to, and not on their degree. Then we compute the clustering coefficient of a node whose neighborhood can be divided in two disjoints sets whose clustering coefficient is known. Remind that in the unipartite version, a bottom node belongs to cliques corresponding to the top nodes it is connected to in the bipartite network.

Lemma III.4: Let $b \in B$ be a node and $T' \subseteq N(b) \subseteq T$ be a set of neighbors of b with degree strictly greater than 2. Then

$$cc(b) \geq \frac{1}{2 \cdot |T'| - 1}$$

Proof: One obtains a lower bound for $cc(b)$ by supposing that all the nodes in T' have no neighbors in common but b. This simpler case brings the lower bound:

$$\begin{aligned} cc(b) &\geq \frac{\sum_{t \in T'} (d(t) - 1)(d(t) - 2)}{\left(\sum_{t \in T'} d(t) - 1\right)\left(\left(\sum_{t \in T'} d(t) - 1\right) - 1\right)} \\ &\geq \frac{\sum_{t \in T'} (d(t) - 1)^2 - \sum_{t \in T'} (d(t) - 1)}{\sum_{t,u \in T'} (d(t) - 1)(d(u) - 1) - \sum_{t \in T'} (d(t) - 1)} \end{aligned}$$

Since $(x - y)^2 \geq 0$ implies $2xy \leq x^2 + y^2$, we have $\sum_{t,u \in T'} (d(t) - 1) \cdot (d(u) - 1) \leq |T'| \cdot \sum_{t \in T'} (d(t) - 1)^2$. Therefore the equation can be bounded by:

$$cc(b) \geq \frac{\sum t \in T'(d(t) - 1)^2 - \sum t \in T'(d(t) - 1)}{|T'| \sum t \in T'(d(t) - 1)^2 - \sum t \in T'(d(t) - 1)}$$

The minimal value of this expression is reached when all the $d(t) - 1$ have the smallest value, i.e. $d(t) = 3$ for all $t \in T'$. We then have:

$$cc(b) \geq \frac{4|T'| - 2|T'|}{4|T'|^2 - 2|T'|} = \frac{1}{2|T'| - 1}$$

This last expression is a lower bound for the clustering coefficient of b in the disjoint case and therefore a lower bound for the general case. ∎

Before introducing the next lemma, we need to define the clustering coefficient of a node b restricted to a subset of neighbors. Given a node b of a network $G' = (B, E')$ and $N' \subseteq N(b)$ a subset of the neighbors of b, the clustering coefficient of b restricted to N' is:

$$cc_{N'}(b) = \frac{|\{(x, y) \in E', x, y \in N'\}|}{\binom{|N'|}{2}},$$

We now give a lower bound for the clustering coefficient of a node in $G' = (B, E')$ whose neighborhood can be divided in two disjoints sub-networks where the clustering coefficient is known for each sub-network:

Lemma III.5: Let b be a node of a network $G' = (B, E')$, and let $S_1, S_2 \subseteq N(b)$ such that $N(b) = S_1 \cup S_2$ and $S_1 \cap S_2 = \emptyset$. Then

$$cc(b) \geq \frac{\alpha \cdot p^2 + 2\beta - 5\beta \cdot p + 3\beta \cdot p^2}{2 - p},$$

where $\alpha = cc_{S_1}(b)$, $\beta = cc_{S_2}(b)$, $p = \frac{|S_1|}{|N(b)|}$, and $\alpha > 0$.

Proof: By definition of the clustering coefficient, we

have:

$$cc(b) = \frac{\alpha\binom{p|N(b)|}{2} + \beta\binom{(1-p)|N(b)|}{2}}{\binom{|N(b)|}{2}}$$
$$= \frac{\alpha p(p|N(b)|-1) + \beta(1-p)((1-p)|N(b)|-1))}{|N(b)|-1}$$
$$= \frac{\alpha p^2(|S_1|-1) + \beta(-1+p)(-|S_1|+p|S_1|+p)}{|S_1|-p}$$

This function is clearly positive, and it is increasing with $|S_1|$ since:

$$\frac{\partial cc(b)}{\partial |S_1|} = \frac{p^2(1-p)(\alpha+\beta)}{(-|S_1|+p)^2} \geq 0$$

Furthermore, we supposed that $\alpha > 0$, which implies that $|S_1| \geq 2$. A lower bound for the clustering coefficient is therefore obtained for $|S_1| = 2$, which gives exactly the claim. ■

We can finally give a lower bound for the clustering coefficient of any node b of $G' = (B, E')$. Among the top nodes connected to b, some have degree 2 and therefore will induce a link in the unipartite version. The extremity of this link may not be linked to any other node in the neighborhood of b and so it does not generate links to be counted for the clustering coefficient of b. On the other hand, b is connected to top nodes of degree greater than 3 which will generate cliques of size greater than 3 and increase the number of links in the neighborhood of b. This gives two disjoints sets of neighbors for which we know the clustering coefficient. We can therefore prove:

Theorem III.6: Let b be a node of $G' = (B, E')$, the unipartite version of $G = (T, B, E)$. Let $T' \subset T$ be the set of the top neighbors of b with degree 2 in G and $T'' \subset T$ the top neighbors with degree strictly greater than 2. Let $B' \subset B$ be the neighborhood of T'' (*i.e.* the bottom nodes which belongs to cliques of size strictly greater than 2 containing b). The fraction of neighbors of b in the unipartite version which belongs to B' will be referred as p. Then

$$cc(b) \geq \frac{p^2}{(2|T''|-1).(2-p)}$$

Proof: First notice that the clustering coefficient of b restricted to the neighbors belonging only to $2-cliques$ (all but B') is 0. Therefore we are going to apply lemma III.5, with a partition of the neighbors in two sets, one having a zero clustering coefficient.

If $|B'| = 0$, then $p = 0$, and obviously $cc(b) = 0$, which fits the formula. On the other hand, if $|B'| > 0$, then by definition it contains at least one link, and therefore has a strictly positive clustering coefficient. This last inequality allows us to apply Lemma III.5 with $S_1 = B'$ which implies that $cc_{B'}(b) \geq \frac{1}{2|T''|-1}$ (from Lemma III.4), and with $S_2 = N(v) \setminus S_1$ (and so $\beta = 0$). ■

To give an intuition on this lower bound, one have to notice that the approximations which lead to the bound concern the probability of the intersection of cliques. The less intersections, the more efficient will be the bound, and therefore, we need a small number of cliques whose sizes are small enough.

Finally, we gave here formal proofs of the fact that our model produces networks having the three main wanted properties. We will see below that they can also be checked experimentally, and that the two methods give results in full agreement.

D. Experimental results

We used two versions of the bipartite model. For the first one, exact top and bottom distribution of degree where used to generate a random network which follows the distributions, therefore the bottom distribution is a power law (Figure 5). For the second one, the bottom distribution is taken to be a Poisson law consistent with the top degree distribution. This will show that it is important to use the bipartite point of view, since both degree distributions are important for the performance of the model.

Both models achieve the goal to get networks with a high clustering coefficient, however only the exact bipartite one fits really well the original complex networks clustering coefficient (see Figure 8), while the Poisson bipartite model generates some networks with a lower clustering coefficient.

networks	C	$C_{Poisson}$	C_{exact}
Actors	0.785	0.443	0.767
Co-authoring	0.638	0.354	0.542
Co-occurrence	0.822	0.099	0.831

Fig. 8. Clustering coefficient of the three main networks obtained by the bipartite model for both exact distributions and bottom Poisson distribution.

The degree distribution of the networks obtained with both models can be compared with the actual networks degree distribution. The Poisson version of the model exhibits some Poisson distribution for the degree whatever the top degree distribution is. On the other hand, the exact bipartite model in which both distributions are respected shows a degree distribution which fits the real one. Figure 9 presents these results. For the three networks we can observe than the real and the exact random bipartite degree distributions are very similar. The only deviation

appears for very small values of the degrees and is hardly visible for both Actors and Co-authoring graphs.

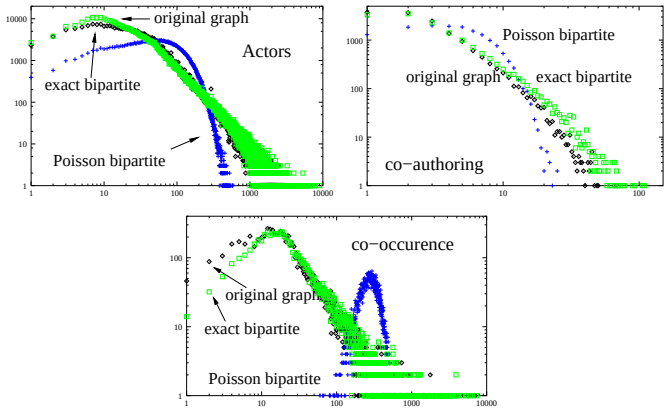

Fig. 9. Degree distribution in the exact and Poisson bipartite model for the Actor, the Co-authoring and the Co-occurrence graphs. Each plot displays degree distribution for the real network, the exact and the Poisson bipartite model.

One can check that the average distance also fits. Therefore our model tends to explain the three main characteristics observed on some complex networks using a simple and observable argument which is their underlying bipartite structure. However, some complex networks of interest, like the Internet for example, do not display such a structure. We deal with these networks in next Section.

IV. Decomposition

In the previous section, we introduced and studied a model based on the underlying bipartite structure of some complex networks, whereas most complex networks do not display such a structure. For example there is no immediate and natural way to see the Internet topology, the Web graph or a Protein graph as bipartite networks. However these networks may contain cliques of high size. We can check experimentally that this is indeed the case: as shown in Figure 11, which we will explain later, there indeed exists large cliques, between 20 and 30 nodes, in the Internet topology, while a Web graph can contain cliques of size 100 and more. A random network with the same size (in terms of nodes and links) contains no clique of size greater than 3 (the existence of cliques of a given size in a random network is known [2] to depend on the connection probability). Such a random network is almost a tree. Therefore these networks display a nontrivial distribution of clique sizes, which makes them very different from random networks. The existence of large cliques in these networks makes it interesting to describe them as bipartite networks as follows: the top nodes are cliques contained in the network we consider, and the bottom nodes are the nodes of the network itself. A clique and a node are linked if the node is contained in the clique.

In this section, we will develop this idea. We will transform a given network into an equivalent bipartite network (recall that different bipartite networks may give the same unipartite network). Once we have this bipartite network, we can reapply the process described in previous section to generate a random bipartite network in order to check that it is similar to the original network, which would show that the properties of the original network can be considered as consequences of the existence of these cliques, encoded in the bipartite network we construct.

A. Decomposition scheme

Our aim here is, given a network $G = (V, E)$, to obtain a set of cliques $C = \{C_i\}$ such that G is the unipartite version of the bipartite network $G' = (C, V, E')$ where $E' = \{\{C_i, v\} | v \in C_i\}$. This problem can also be viewed as follows: we look for a set of cliques such that the links in E are exactly the links in the cliques. It is known as the *clique covering problem* [37], [38].

A trivial solution is given by $C = E$: each clique covers exactly one link of the network. However, our aim is to obtain a set of cliques C such that the bipartite network G' will have properties similar to the ones observed for natural bipartite networks: large cliques should be discovered and the number of cliques should be linear in the size of the network.

Minimizing the number of cliques leads to the *minimal clique covering problem* which is known to be NP-complete [37], [38]. Computing maximal cliques of a network is also NP-complete [39], [40]. However, some heuristics make it possible to compute them if the network is not too large. In our case, we use the following remarks. Recall that we denote by $N(u)$ and $N(u, v)$ the sets $\{v \in V | \{u, v\} \in E\}$ and $N(u) \cap N(v)$ respectively. First notice that a largest clique containing $\{u, v\}$ in G is also a largest clique containing $\{u, v\}$ in the sub-network of G induced by $N(u, v) \cup \{u, v\}$. Moreover, if we denote by $\mathcal{C}$ the largest clique in the sub-network of G induced by $N(u, v)$, then $\mathcal{C} \cup \{u, v\}$ is the clique we are looking for. See Figure 10.

In complex networks, we observed that the sub-networks induced by $N(u, v)$ for all links $\{u, v\}$ are very dense and very small , which makes it possible to compute the following clique covering of these complex networks: for each link $\{u, v\}$ in G we take the largest clique containing it (if there are more than one, we choose one largest clique at random). We obtain this way a number of cliques bounded by the number of links in the network. Moreover, the obtained clique size distributions

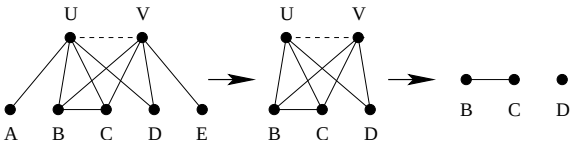

Fig. 10. Given a network $G = (V, E)$, we are looking for a largest clique containing the link $\{u, v\}$. This clique is necessarily contained in the sub-network induced by $N(u, v) \cup \{u, v\} = \{b, c, d, u, v\}$. It is actually sufficient to compute the largest clique $\mathcal{C}$ in the sub-network induced by $N(u, v) = \{b, c, d\}$ since the clique we are looking for is nothing but $\mathcal{C} \cup \{u, v\}$ which, in our case, gives $\{u, v, b, c\}$

(Figure 11) show that this scheme achieves our goal.

B. Experimental Results

Given the decomposition scheme described above, we can now transform any complex network into a bipartite network. Figures 11 and 12 show the top and bottom distribution obtained for Internet, Web and Protein networks.

These distributions exhibit a structure similar to the one observed on natural bipartite networks. First notice that bottom degree distribution fit very well power laws. Two of the top degree distribution, the ones of Internet and Protein networks clearly follow a Poisson law, while the one of the Web graph is more heavy tailed. Notice that these distribution exhibit some surprising behaviors, like the presence of many cliques of size greater than 20 in some Internet topologies, which even contain 33 cliques of size 154. Some of these behaviors are discussed in Section V.

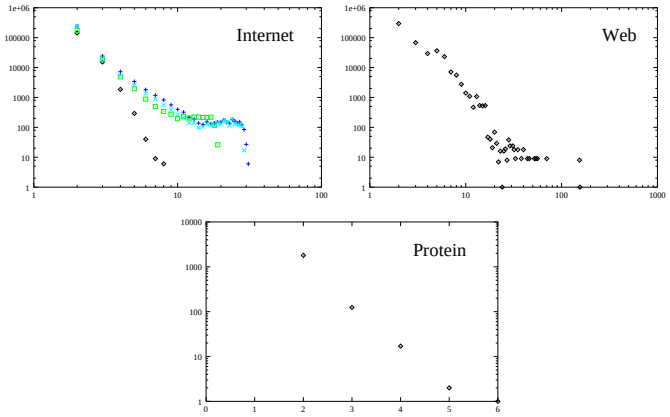

Fig. 11. Clique size (top degree) distribution obtained from Internet, Web and Protein graphs.

C. Recomposing a network

A shown in Section III, our model is relevant in the sense that the properties of bipartite complex networks can be viewed as consequences of their natural bipartite

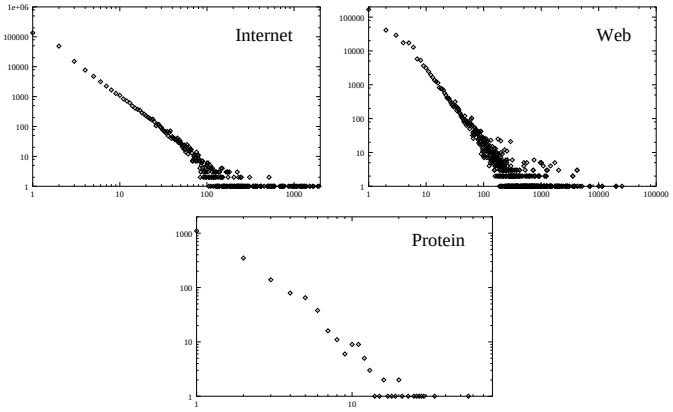

Fig. 12. Bottom degree distribution distribution obtained from Internet, Web and Protein graphs.

structure. We are now wondering if it is also the case for complex networks which do not have such a structure naturally. This would give an indication of the fact that the existence of non-trivial cliques in complex networks with some nodes belonging to many of them is of prime importance for the similarities between all complex networks. Consequently our final step consists in generating a random bipartite network with the model, using the distributions obtained by the decomposition process, and then compare its unipartite version with the original network.

Figure 13 shows that the degree distributions of the recomposed networks fit very well the original data: the generated networks clearly exhibit a power law degree distribution due to the bottom degree distribution. Notice however the deviations for both Internet and Web graphs resulting in a peak in the distributions (around 20 for Internet and 150 for Web graph). These perturbations come from the existence of very large cliques in these networks (see Figure 12). Despite this, the degree distribution is very well captured even for networks which do not display a natural bipartite structure.

The clustering coefficient obtained for recomposed networks is shown in Figure 14 together with the original one. The values are very similar but tends to be greater for the recomposed networks, in particular for the recomposed Internet topology (one order of magnitude higher than the original one). This is due to the existence of many large cliques in the original network (see Figure 11) which tend to be disseminated in the recomposed network and therefore increase the overall clustering coefficient. One can wonder whether such cliques really exists in the Internet topology or not, and the effects they have on the behavior of the model. This is discussed in Section V.

Finally, these results show that the properties – high clustering, low average distance and power law degree distribution – observed on complex networks can there-

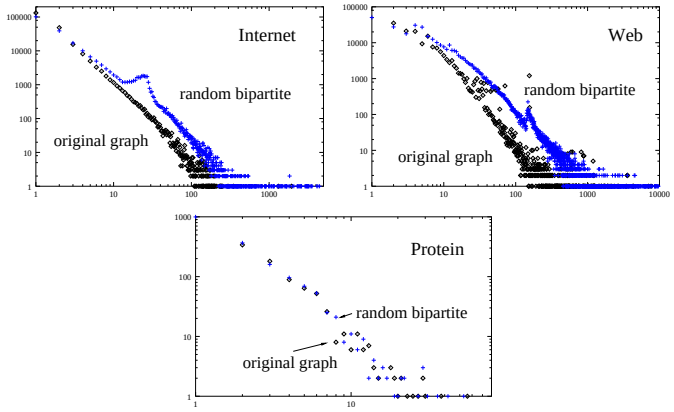

Fig. 13. Degree distribution for Internet, Web and Protein graphs. For each, original distribution and distribution for the networks obtained using the decomposition scheme and the bipartite model.

Networks	$C_{initial}$	C_{model}
Internet	0.060	0.456
Protein	0.153	0.187
Web	0.466	0.663

Fig. 14. Original clustering coefficient of three complex networks which do not have an immediate bipartite structure, together with the clustering coefficient of the corresponding recomposed networks.

fore be understood as the consequence of the underlying bipartite structure, either natural for networks constructed this way, or due to a nontrivial distribution of cliques size and the existence of nodes which belongs to many cliques.

V. Conclusion and discussion

In this paper, we have proposed a complex network model which achieves the following challenges:

- it has the three main wanted properties (logarithmic average distance, high clustering and power law degree distribution),
- it is based on a *realistic* construction process representative of what happens for some real complex networks, and
- its definition is simple enough to make it possible to give some intuition and some proofs of its properties.

Whereas many models have already been introduced, this one is the first which reaches both goals at the same time. In this sense, it may be considered as a new step towards the realistic modeling of complex networks. Moreover, it is very simple to obtain networks using this model (we provide a network generator at [41]), which makes it highly suitable for simulation purpose.

The model is based on the remark that some complex networks have an underlying bipartite structure which can be seen as responsible for their main properties. Despite the fact that most complex networks do not have this natural bipartite structure, we show that they actually can be decomposed into cliques which make such a structure emerge. This shows that the main properties of complex networks can be viewed as consequences of this bipartite structure, and that the model captures a very general behavior of complex systems.

Another contribution of our work is the computation of new kinds of statistics on complex networks, namely the clique size distribution, the way neighborhoods intersect, and the way cliques overlap. Obtaining new statistics on complex networks is important in order to understand the details of their structure, the relevance of various models, and many other problems. We discuss these aspects below.

Clique size distribution. Notice that, whereas it seems quite natural to find large cliques in the Web graph (for example, a set of pages containing a menu in which each item points to a page in the set induces a large clique) one might be disappointed by the fact that quite a large number of quite large cliques (typically several dozens of cliques of size greater than 20) appear in Internet topologies. Is it really possible that 30 routers on the Internet are all pairwise physically connected? Clearly not. On the contrary, this statistics should be understood as an evidence of a distortion induced by the measurement method. The way one explores the Internet topology (mainly using `traceroute` and BGP tables) give biased views of some special structures, like tunnels or ATM sub-networks, which may appear like cliques. In this context, our statistics can be viewed as a way to identify such artefacts.

Neighborhood intersection. During our work, we had to face the NP-hard problem of computing cliques on very large networks. At first, we believed that this was an impossible goal and we planned to use approximation algorithms. However, it appeared that we could use the specific properties of complex networks (namely their clustering coefficient) to manage exact computations efficiently. Indeed, the maximal clique containing a given link $\{u, v\}$ is included in $\{u, v\} \cup (N(u) \cap N(v))$. This sub-network is quite small and it is very dense (in many cases it is almost a clique). This is why computing cliques on these sub-networks to obtain the cliques of the complex networks we consider is efficient. This shows that statistical properties of complex networks have a great impact on how we can algorithmically manage them, which should be taken into account when one wants to write algorithms for complex networks.

Finally, there are many directions in which this work may be extended. There is still much to do in the modeling of complex networks in order to understand them pre-

cisely. The correlations between the degrees in a bipartite network and in its unipartite version should be further investigated. Indeed, it seems that these correlations are of different kinds, depending of the complex network in observation. Likewise, the notion of clustering is only partially captured by the notion of *clustering coefficient*. It is certainly important to study it more precisely, by considering the clustering coefficient as a function of the degree of the nodes, for example [42]. Other parameters, like a clustering coefficient at a given distance, or a clustering attached to the links, may also be relevant.

REFERENCES

[1] P. Erdos and A. Renyi, "On random graphs i," *Publ. Math. Debrecen*, vol. 6, pp. 290–297, 1959.

[2] Bela Bollobas, *Random Graphs*, Academic Press, 1985.

[3] R. Pastor-Satorras and A. Vespignani, "Epidemic spreading in scale-free networks," *Phys. Rev. Lett.*, vol. 86, pp. 3200–3203, 2001.

[4] M. E. J. Newman, "The spread of epidemic disease on networks," *Phys. Rev. E*, vol. 66, 2002.

[5] Hawoong Jeong Réka Albert and Albert-László Barabási., "Error and attack tolerance in complex networks.," *Nature*, vol. 406, pp. 378–382, 2000.

[6] Adilson E. Motter and Ying-Cheng Lai., "Cascade-based attacks on complex networks.," *Physical Review E 66*, 2002.

[7] Toby Walsh, "Search in a small world," in *IJCAI*, 1999, pp. 1172–1177.

[8] Beom Jun Kim, Chang No Yoon, Seung Kee Han, and Hawoong Jeong., "Path finding strategies in scale-free networks.," *Phys. Rev. E 65, 027103.*, 2002.

[9] Damien Magoni and Jean-Jacques Pansiot., "Influence of network topology on protocol simulation.," in *ICN'01 - 1st IEEE International Conference on Networking*, July 9-13, 2001, vol. Lecture Notes in Computer Science, pp. 762–770.

[10] R. Albert and A. Barabasi, "Statistical mechanics of complex networks," *Reviews of Modern Physics 74, 47*, 2002.

[11] Steven H. Strogatz., "Exploring complex networks.," *Nature 410*, March 2001.

[12] S.N. Dorogovtsev and J.F.F. Mendes, "Evolution of networks," *Adv. Phys. 51, 1079-1187*, 2002.

[13] Internet Maps from Mercator, "http://www.isi.edu/div7/scan/mercator/maps.html," .

[14] Q. Chen, H. Chang, R. Govindan, S. Jamin, S. Shenker, and W. Willinger, "The origin of power laws in internet topologies revisited," in *INFOCOM*, 2002.

[15] Self-Organized Networks Database, "http://www.nd.edu/˜networks/database/index.html," .

[16] The Internet Movie Database, "http://www.imdb.com/," .

[17] Today New International Version, "http://www.tniv.info/bible/," .

[18] arXiv.org e Print archive, "http://arxiv.org/," .

[19] H. Jeong, B. Tombor, R. Albert, Z. Oltvai, and A. Barabasi, "The large-scale organization of metabolic networks," *Nature, 407, 651*, 2000.

[20] C. Moore and M. E. Newman, "Epidemics and percolation in small-worlds networks," *Phys. Rev. E*, vol. 61, pp. 5678–5682, 2000.

[21] D. Watts and S. Strogatz, "Collective dynamics of small-world networks," *Nature*, vol. 393, pp. 440–442, 1998.

[22] S.N. Dorogovtsev and J.F.F. Mende, "Exactly solvable small-world network," *Euro. phys. Lett.*, vol. 50 (1), pp. 1–7, 2000.

[23] Albert-Laszlo Barabasi and Reka Albert, "Emergence of scaling in random networks," *Science*, vol. 286, pp. 509–512, 1999.

[24] S. Dorogovtsev, J. Mendes, and A. Samukhin, "Structure of growing networks with preferential linking," *Phys. Rev. Lett. 85*, pp. 4633–4636, 2000.

[25] F. Comellas, G. Fertin, and A. Raspaud, "Vertex labeling and routing in recursive clique-trees, a new family of small-world scale-free graphs," in *Sirocco 2003 - The 10th Int. Colloquium on Structural Information and Communication Complexity*, pp. 73–87.

[26] Albert-Laszlo Barabasi, Erzsebet Ravasz, and Tamas Vicsek, "Deterministic scale-free networks," *Physica A 299, (3-4)*, pp. 559–564, 2001.

[27] B. M. Waxman, "Routing of multipoint connections," *IEEE Journal of Selected Areas in Communications*, pp. 1617–1622, 1988.

[28] Ellen W. Zegura, Kenneth L. Calvert, and Michael J. Donahoo, "A quantitative comparison of graph-based models for Internet topology," *IEEE/ACM Transactions on Networking*, vol. 5, no. 6, pp. 770–783, 1997.

[29] T. Bu and D. Towsley, "On distinguishing between internet power law topology generators," in *INFOCOM*, 2002.

[30] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos, "On power-law relationships of the internet topology," in *SIGCOMM*, 1999, pp. 251–262.

[31] William Aiello, Fan R. K. Chung, and Linyuan Lu, "A random graph model for massive graphs," in *ACM Symposium on Theory of Computing*, 2000, pp. 171–180.

[32] Alberto Medina, Ibrahim Matta, and John Byers, "On the origin of power laws in internet topologies," in *ACM Computer Communication Review, 30(2)*, april, 2000.

[33] C. Jin, Q. Chen, and S. Jamin, "Inet: Internet topology generator," Tech. Rep. CSE-TR-443-00, Department of EECS, University of Michigan, 2000.

[34] Alex Fabrikant and et al., "Heuristically optimized trade-offs: A new paradigm for power laws in the internet, icalp 2002," .

[35] D. Magoni and J.-J. Pansiot, "Internet topology modeler based on map sampling," in *Proceedings of ISCC'02, IEEE Symposium on Computers and Communications*, Italy, July 2002.

[36] Linyuan Lu, "The diameter of random massive graphs," in *12th Ann. Symp. on Discrete Algorithms (SODA)*, ACM-SIAM, Ed., 2001, pp. 912–921.

[37] J.Orlin, "Contentment in graph theory: Covering graphs with cliques," *Indigationes Mathematicae*, vol. 80, pp. 406–424, 1977.

[38] Sylvia D. Monson, N. J. Pullman, and Rolf. Rees, "A survey of clique and biclique coverings and factorizations of (0,1)-matrices.," *Bull. Inst. Combin. Appl.*, vol. 14, pp. 17–86, 1995.

[39] I. Bomze, M. Budinich, P. Pardalos, and M. Pelillo, "The maximum clique problem," in *Handbook of Combinatorial Optimization*, D.-Z. Du and P. M. Pardalos, Eds., vol. 4. Kluwer Academic Publishers, Boston, MA, 1999.

[40] J. Abello, P. Pardalos, and M. Resende, "On maximum clique problems in very large graphs," *External Memory Algorithms, DIMACS Series, AMS*, 1999.

[41] Source code for the generator, "http://www.liafa.jussieu.fr/˜guillaume/programs/," .

[42] A.L. Barabasi, Z. Deszo, E. Ravasz, S. H. Yook, and Z. Oltvai, "Scale-free and hierarchical structures in complex networks," in *Sitges Proceedings on Complex Networks*, 2004.