Bipartite Graphs as Models of Complex Networks

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

Abstract

It appeared recently that the classical random graph model used to represent real-world complex networks does not capture their main properties. Since then, various attempts have been made to provide accurate models. We study here the first model which achieves the following challenges: it produces graphs which have the three main wanted properties (clustering, degree distribution, average distance), it is based on some real-world observations, and it is sufficiently simple to make it possible to prove its main properties. This model consists in sampling a random bipartite graph with prescribed degree distribution. Indeed, we show that any complex network can be viewed as a bipartite graph with some specific characteristics, and that its main properties can be viewed as consequences of this underlying structure. We also propose a growing model based on this observation.

Introduction.

When one wants to model a real-world object (in the sense of producing an artificial object similar to the real one), one first has to get some information on its properties, generally using a measurement procedure and an analysis of the result of this measure. There are then basically two ways to propose a model.

First, one may consider a set of observed properties as essential, and then sample randomly objects among the ones which have these properties. One obtains this way a *typical* object with the properties in concern. It is then possible to determine if the retained set of properties is sufficient (do the random objects produced by the model fit well the real one?) and to study the expected behavior of the object of interest. The relevance of the set of properties is generally checked using other known properties or behaviors of the object.

The other modeling approach is to define a construction process inspired from the way the object is *really* constructed. This construction process is generally iterated from an initial state, and eventually leads to an appropriate object. The analysis then concerns the properties induced by the construction process: do they fit real-world properties?

The first method is in general more suitable for analysis, and more rigorous, but it may be very difficult to sample a random object in a given class. On the opposite, the second approach generally gives a simple sampling scheme and has the advantage of producing *evolving* objects. But the construction process may induce some properties which do not correspond to any reality, and is in general difficult to analyze.

It has been shown recently that most real-world complex networks have some essential properties in common. These properties are not captured by the model generally used before this discovery, although they play a central role in many contexts like the robustness of the Internet [6, 19, 20, 15, 49], the spread of viruses or rumors over the Internet, the Web or other social networks [48, 52, 58], as well as the performance of protocols and algorithms [36, 41, 63].

This is why, in the last few years, a strong effort has been put in the realistic modeling of complex networks, both in computer science, mathematics and physics, and much progress has been accomplished in this field. Some models achieve the aim of producing graphs which capture some, but not all of the main properties of real-world complex networks. Some models obtain all the wanted properties but rely on artificial methods which give unrealistic graphs (trees, graphs with uniform degrees, etc). Others rely on construction processes which may induce some hidden properties, or are too difficult to analyze.

In this paper, we propose the random bipartite graph model as a general model for complex networks. It has all the advantages we have just cited, without the drawbacks. It produces graphs with all the wanted properties. It relies on real-world observations and gives realistic graphs. Finally, it is simple enough to make it possible to prove its main properties.

We will first present an overview of the context in which our work lies. In particular, we use some ideas introduced in previous papers, which we need to describe precisely. Then we show how *all* complex networks may be described as bipartite structures. After this, we present the random bipartite model and analyze it to show that the main properties of complex networks are somehow a consequence of their underlying bipartite structure. We also present a growing bipartite model based on the same ideas. Finally we discuss the advantages and limitations of these models.

1 Context.

Throughout our presentation, we will use a representative set of complex networks which have received much attention and span quite well the variety of contexts in which complex networks appear. These complex networks can be divided in three main classes, namely social networks, technological networks and biological ones. The set consists of:

- ***Internet.*** The interconnection of routers (or AS) on the Internet can be modeled by graphs where the nodes are routers (or groups of routers) linked with physical links. We use several graphs from [16, 32, 33]

- ***Web.*** The hyperlinks between Web pages give a natural graph structure to the World Wide Web [14, 37]. We will use here the Notre Dame Web graph from [5, 23].

- ***Cooccurrence.*** When one considers a book, or the queries to a search engine, or a chat on an interactive system for instance, one can construct a co-occurrence graph by linking two words if they appear in the same sentence or query [31]. Here, we will use a version of the Bible [62].

- ***Actors.*** In this social network, two actors are connected if they have played together in a movie. This graph is widely studied for many reasons: it is very large, well representative of social networks, evolving with each new movie produced, and easily available through the Internet Movie Database [24, 64].

- ***Coauthoring.*** Another way to link people is according to their scientific publications: two scientists are linked if they have signed a paper together [50, 51, 56]. We will use such a graph obtained from the Los Alamos preprint archive [7].

- ***Proteins.*** In [35] the authors link together two proteins of a given biological system if they influence each other. We will consider this example too, using graphs from [23].

Many other complex networks have been studied. Refer to [4, 26, 54] for a more descriptive list of networks and corresponding references. All these networks have some properties in common which have been discovered quite recently and have concentrated a large attention in various communities. Hereafter we present the properties in concern and some recent efforts in the modeling of these properties.

1.1 Statistical properties

Most real-world complex networks have a number of edges m which scales linearly with the number of vertices n: $m \sim k \cdot n$ where k is the average degree (which does not depend on the size of the graph). Therefore, these networks have a low density (going to 0 when n grows), the density being defined as the number of existing edges over the number of edges that could exist.

Three other properties received recently much attention due to the fact that they have unexpected behaviors in real-world complex networks: the average distance between vertices, the clustering and the degree distribution.

The distance between two vertices, defined as the number of edges on a shortest path between these vertices, is low on average. It is a well known property on social networks since the work of Stanley Milgram [43] and the notion of "six degrees of separation". However it appeared more recently that all complex networks have a low average distance which typically scales like the logarithm of the size of the graph. It has been shown that this is actually true for any graph which contains some resaonable amount of randomness. Actually, under reasonable assumptions, the average distance in random graphs scales even slower than the logarithm [1] of their size [12, 18, 21, 28, 39, 55, 56].

The local clustering [64] is defined for each vertex of degree at least 2 as the proportion of edges between its neighbors: $c(u) = \frac{|\{(x,y), x,y \in N(u)\}|}{\binom{d(u)}{2}}$, where $d(u)$ is the degree of vertex u and $N(u)$ is the set of neighbors of u. The global clustering is simply the average over all individual values. Another definition (a global one) set the clustering of a graph to be the ratio of the number of triangles (three vertices all connected) over the number of connected triples (three vertices with at least two edges) [56]: $c_g(u) = \frac{3 \cdot |\triangle|}{|\wedge|}$. Even if both definition are not strictly equivalent, one can understand the clustering as a measure of the local density of a graph: it is the probability that two neighbors of a vertex are connected together. Hereafter we are going to use the first definition which is more widely accepted. Although most graphs

[1] Typically like the logarithm divided by the logarithm of the logarithm.

	Internet	Web	Actors	Co-auth	Co-occur	Protein
n	75885	325729	392340	16401	9297	2113
m	357317	1090108	15038083	29552	392066	2203
density	1.2e-4	2.1e-5	1.9e-4	2.2e-4	9.1e-3	9.9e-4
α	2.5	2.3	2.2	2.4	1.8	2.4
c	0.171	0.466	0.785	0.638	0.822	0.153
d	5.80	7	3.6	7.18	2.13	6.74

Table 1: The main statistics for the complex networks we use in this paper. For each network, we give its number of vertices n, its number of links m, its density, the value of the exponent α of the power law that fits best its degree distribution, its clustering c, and its average distance d.

have a very low clustering (inversely proportional to the size of the graph if $m \sim k \cdot n$), all real-world complex networks have a high clustering which seems to be independent of the size of the network.

Finally, the degree distribution which is, for each k, the probability p_k that a randomly chosen vertex has degree k, is completely different from what was expected. Indeed for almost all real-world complex networks, the degree distribution follows a power law: $p_k \sim k^{-\alpha}$, while one would have expected an exponential decrease (Poisson-like distributions). The exponent α of the power law is generally between 2 and 3. Such a distribution means that although most vertices have a small degree, the number of vertices with degree k decays only polynomially with k, and therefore there is a significant number of vertices with high degree.

The main properties of the real-world complex networks we use in this paper are given in Table 1. Notice that, as announced, all these real-world complex networks have a very low density, a low average distance, a power law distribution of degrees and a high clustering.

The similarity of these networks concerning unexpected properties has led to the study of other properties. The simplest one concerns the degree-degree correlation: what is the average degree of the neighbors of a vertex of degree k. Three main behaviors are expected, either high-degree vertices tend to connect to high-degree vertices, or to low-degree vertices, or to any nodes. These behaviors can be observed using the the slope (increasing, decreasing or constant) of the plot which relates the average degree of the neighbors of nodes of degree k, to k [10, 60, 61], or with a single parameter (assortativity coefficient), which may be positive (assortative networks), negative (dissortative networks) or null (neutral networks) [53]. Most social networks are assortative (similar vertices are connected) while technological or biological are generally dissortative.

One may also correlate the clustering and the degree by computing the average clustering of vertices having a given degree. This also defines assortative (high degree yields high clustering), neutral or dissortative networks.

Finally, other properties have been studied, such as the centrality (how many shortest paths contain a given vertex) [51], the distribution of eigenvalues of the adjacency matrix [30, 42], etc. All these statistical properties are used to describe a given complex network and to study the similarities and differences between several complex networks. They give precise insight on what one may expect when considering a complex network having a set of properties.

1.2 Modeling complex networks

The basic model for complex networks is the Erdös-Rényi (ER) random graph model [12, 29]. In a random graph with n vertices, each of the $\frac{n\cdot(n-1)}{2}$ possible edges exists with a given probability p (this model is know as $\mathcal{G}_{n,p}$). In an equivalent way when n tends to infinity [12, 29], one may construct such a random graph from n vertices by choosing $m = p \cdot \frac{n\cdot(n-1)}{2}$ edges at random ($\mathcal{G}_{n,m}$ model).

In such a graph, it is known that the average distance scales with the logarithm of n [12]. Moreover, the clustering is equal to the connection probability p since each pair of vertices is connected with the same probability independently of the fact that they are both linked to a same vertex. If $m \sim k \cdot n$ as in real-world complex networks, this means that the clustering scales as n^{-1} and therefore tends to 0 when n grows. Finally, the degree distribution follows a Poisson law $p_k \sim e^{-\lambda}\frac{\lambda^k}{k!}$ [12], which implies in particular that the number of vertices with degree k decays very rapidly around the average degree, and therefore all vertices have nearly the same degree.

Therefore, although this model can be considered as relevant concerning the average distance, it misses the two other main properties of real-world complex networks. In particular, the degree distributions are qualitatively different.

It is however possible to sample uniformly a random graph with a given degree distribution (in particular a power law) [11, 40, 45, 46] using the Molloy and Reed[2] (MR) model: for each vertex, draw its degree at random according to the given distribution, create as many connection points as its

[2] Despite it has been introduced in [9] and studied in [11], this model is commonly refferred to as the *Molloy and Reed* model since these authors made it popular in their contributions [45, 46]. We will follow this convention here.

degree and finally connect pairs of connection points at random[3] (Figure 1). Notice that the degree distribution can be explicitly described (four nodes of degree 1 and one node of degree 4 for instance) or implicitly defined (power law with exponent 2.2 for instance).

The power law graphs obtained this way have an average distance which scales slower that the logarithm[4] of their size [18, 21, 28, 39, 55, 56]. Moreover, the fact that vertices are linked together purely at random (only their number of edges is given) makes it possible to study the properties of the obtained graphs, and it indeed seems that it captures some of the most important behaviors of complex networks [2, 18, 39, 54, 58]. However, under reasonable assumptions on the degree distribution, the clustering of these graphs tends to zero when n grows [54].

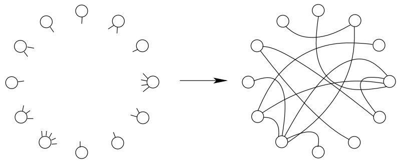

Figure 1: MR model with prescribed degree distribution. Pairs of connection points are randomly chosen to create the links, while the number of connection points of every node follows a given distribution.

Both the purely random graph model and the one with prescribed degree distribution belong to the class of models which sample uniformly at random a graph among the ones with a given set of properties (number of vertices and average degree for the first, number of vertices and degree distribution for the second). This approach could in principle be continued, and sampling a random graph among the ones having a given number of nodes, a given degree distribution *and* a given clustering would certainly be an excellent model. However, until now, there is no known method to sample such a graph, and the problem seems difficult.

On the other hand, a large variety of models based on the iteration of a construction process *inspired* from the way complex networks grow in reality have been introduced.

[3] Note that this algorithm may induce multiple links and loops. Since, under reasonable assumptions, their number goes to 0 when the graph grows, they are usually neglected in complex network studies. We will follow this convention here. One may also use techniques to avoid them, see for instance [38, 44]i, but this is out of the scope of this paper.

[4] Typically like the logarithm divided by the logarithm of the logarithm.

The first generic model of real-world complex networks, and one of the most famous, has been introduced in 1998 by Watts and Strogatz (WS) [64]. One starts with a ring of n vertices in which each vertex is connected to its k nearest neighbors, for a given k. Then, each edge is rewired with a given probability p by choosing randomly a new extremity (Figure 2).

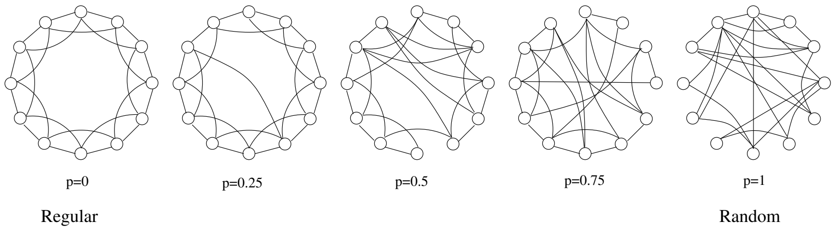

Figure 2: WS model: between regularity and randomness the graph has both low average distance and high clustering.

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 edges are rewired and so the graph is almost a ring. Notice however that, since each vertex is connected to its nearest neighbors, these neighbors are linked together and so the clustering is high. On the other hand, if p is high, then almost all the edges are rewired, and so the graph is similar to a random graph: the average distance is low and so is the clustering. For medium values of p, the graph has both a small average distance and a high clustering, which corresponds to two general properties of complex networks. Moreover, some properties of this model can be formally studied (see for instance [25]), but the degree distribution of the obtained graphs does not follow a power law.

Another important step was done when Albert and Barabási (AB) introduced their model based on *preferential attachment* [3, 27]. The idea can be well understood if we think about the way new Web pages connect to existing ones. Intuitively, when one creates a new Web page, one more likely connects it to a well known one rather than to 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 vertices arrive one by one in a graph and choose their neighbors with a probability proportional to the degree of these neighbors. This model has been studied intensively and is now well known (see [4] for a survey of its properties). For instance, the degree distribution follows a power law with exponent 3. The average distance of such a graph is logarithmic in the number of vertices, and the clustering is low, going to 0 when the number of vertices grows. Despite this last point, this model has received much attention, in particular because it defines *growing* graphs. We will see in Section 3 that the preferential attachment principle can be used to define a growing *bipartite* model with interesting properties.

Both the WS model and the AB one have been introduced to model generic behavior of complex networks. However, they both fail in producing graphs having each of the three properties we cited. The WS model gives a possible explanation for the high clustering of complex networks which is the locality of the links. On the other hand the AB model gives an explanation to the power law degree distribution with the "preferential attachment" principle. Both concepts have been widely used as building blocks for more complex models.

One of them is the Dorogovstev and Mendes (DM) model which generates highly clusterised graphs with a power law degree distribution [26]. This model is very similar to the AB model: for each newly created vertex, an edge is chosen at random and the new vertex is connected to both extremities of the edge. Since high-degree vertices have more edges, they are more likely to be chosen. The preferential attachment is therefore hidden in this model. Moreover, each new vertex is linked to two previously connected vertices, which creates a triangle and induces high clustering. However the parameters of this model cannot be tuned and it has some unexpected properties (for instance, there is no node of degree 1 and it produces *planar* graphs[5]). Therefore we are not going to use it hereafter.

Some deterministic models, which we do not detail here, have also been introduced [8, 22] which produce the wanted properties and are suitable for analysis. However, they cannot be considered as realistic and the obtained graphs have specific properties which make them very different from real-world complex networks.

Many other attempts have been made to reach the goal of obtaining growing models which give graphs having each of the three main properties we have cited. Most of them are described in [4, 26, 59]. However, all these models fail to give an intuitive, realistic and simple interpretation of the causes of the observed properties. Even if these models are based on the simulation of a construction process *inspired* from reality, which makes them more realistic, the drawback comes from the difficulty to analyze them in general. Finally, as already stressed, the construction process may induce

[5] A graph is planar iff it can be embedded in the plane so that no edges intersect.

	Internet	Web	Actors	Co-auth	Co-occur	Protein
n	75885	325729	392340	16401	9297	2113
m	357317	1090108	15038083	29552	392066	2203
c	0.171	0.466	0.785	0.638	0.822	0.153
c_{ER}	0.0001	0.00002	0.0002	0.0002	0.009	0.001
c_{MR}	0.0694	0.017	0.0057	0.001	0.26	0.007
c_{AB}	0.0024	0.0005	0.0015	0.003	0.028	0
c_{WS}	0.171	0.461	0.74 (*)	0.523 (*)	0.74 (*)	0.06 (*)
d	5.80	7	3.6	7.18	2.13	6.74
d_{ER}	5.25	5.47	2.97	7.57	2.06	10.4
d_{MR}	3.25	4.48	2.95	5.77	2.36	5.73
d_{AB}	4.15	5.1	2.93	5.5	2.38	8.15
d_{WS}	5.90	11.23	2559 (*)	2269 (*)	55.6 (*)	509 (*)

Table 2: Performance of the main generic models for complex networks. For each network, we give its number of vertices n, its number of links m, its clustering c, and its average distance d. Moreover, we give the values of these parameters for typical graphs with the same number of vertices and edges obtained with commonly used models: the ER model (c_{ER} and d_{ER}), the MR model (c_{MR} and d_{MR}), the AB model (c_{AB} and d_{AB}), and the WS model (c_{WS} and d_{WS}). The cases pointed by a star (*), the real clustering is too large to be obtained with the WS model. Therefore we used in these cases the parameters inducing the maximal clustering, which yields very large average distances.

	density	average distance	degree dist	clustering
ER	OK	OK	NO	NO
MR	OK	OK	OK	NO
AB	OK	OK	OK	NO
WS	OK	OK	NO	OK

Table 3: Properties captured by the main current models.

some unwanted properties on the obtained graphs.

Table 2 shows the performances obtained with the main models we cited in our practical cases. Let us insist on the fact that the models seek *qualitative* properties (non negligible clustering, power law degree distribution, etc). Their aim is not to produce graphs with exactly given values for these properties. However, even with this in mind, the graphs obtained using these models are significantly different from real-world ones concerning at least one of these three points.

1.3 Current state of the art

This overview of complex networks analysis and modeling shows that although much progress has been accomplished, we still do not have any realistic model to produce graphs with the three main properties of real-world complex networks: small average distance, high clustering and power law degree distribution (see Table 3 for a synthetic view of the properties captured by the main models). The random sampling of graphs among the ones having a set of properties seems natural and promising. It leads to rigorous studies and strong insight on how real-world complex networks behave and on the influence of their specific properties. However, there is no known method to sample a graph with the three wanted properties: the clustering misses. On the other hand, the models based on the iteration of a construction process have inherent disadvantages, like the complexity of their analysis or hidden unwanted properties possibly induced by the construction process. Moreover, until now, no realistic model of this kind has been introduced which has the three wanted properties *and* has been rigorously studied. However, these models have the important advantage of producing *growing* graphs, *i.e.* graphs which grow during time. Notice also that the study of the way real-world complex networks are constructed is highly relevant and is a challenge in itself.

In this paper, we propose a solution to the random sampling of graphs which have all the three wanted properties. To achieve this, we focus on another property of *all* real-world complex networks, namely their underlying bipartite structure (Section 2). We then propose two models: the random sampling of bipartite graphs with prescribed degree distributions, and the growing bipartite model with preferential attachment (Section 3). Indeed, as shown in Sections 4 and 5, respectively formally and experimentally, these models induce the three wanted properties. This means that they can be viewed as consequences of the underlying bipartite structure of all complex networks, which is our main contribution.

2 Complex networks as bipartite graphs

A bipartite graph is a triple $G = (\top, \bot, E)$ where $\top$ and $\bot$ are two disjoint sets of vertices, respectively the top and bottom vertices, and $E \subseteq \top \times \bot$ is the set of edges. The difference with classical graphs lies in the fact that edges exist only between top vertices and bottom vertices.

Two degree distributions can naturally be associated to such a graph, namely the *top degree distribution*: $\top_k = \frac{|\{t \in \top : d(t)=k\}|}{|\top|}$ and the *bottom degree distribution*: $\bot_k = \frac{|\{t \in \bot : d(t)=k\}|}{|\bot|}$. These two distributions play a central role in the following.

Natural bipartite structures

As already noticed for instance in [55, 34], some complex networks display a natural bipartite structure. Among our examples, one can view *Actors* (two actors are linked if they are part of a same cast) as a bipartite graph where $\top$ is the set of movies, $\bot$ is the set of actors, and each actor is linked to the movies he/she played in. *Coauthoring* can also be viewed this way with $\top$ being the set of papers and $\bot$ being the set of authors, each author being linked to the papers he/she (co-)signed. Likewise, in *Cooccurrence* one can link each sentence to the words it contains.

Given a bipartite graph $G = (\top, \bot, E)$, one can easily obtain its classical version, also called $\bot$-projection, defined as $G' = (\bot, E')$ where $\{u, v\}$ is in E' if u and v are both connected to a same (top) node in G. See Figure 3 for an example. From the bipartite versions of *Actors*, *Coauthoring* and *Cooccurrence* graphs, one can then recover their classical versions. In the $\bot$-projection of a bipartite graph, each top vertex induces a clique (complete subgraph) between the bottom vertices to which it is linked: all actors of a given movie have played together therefore they must be all linked.

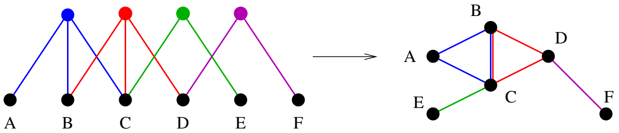

Figure 3: A bipartite network and its $\perp$-projection. Notice that the link $\{B, C\}$ is obtained twice since B and C have two neighbors in common in the bipartite network.

However, given the $\perp$-projection of a bipartite graph, it is in general not possible to recover the bipartite graph from which it has been obtained in an unique way. Similarly if a graph is not naturally bipartite there may exist many bipartite versions of it.

Recovering a bipartite structure

For the sake of completeness, we now recall and detail the decomposition scheme we proposed in [34], which produces a bipartite graph from any given graph, such that the latter is be the $\perp$-projection of the obtained bipartite graph. The aim of this scheme is that the obtained bipartite graph should have properties similar to the ones met in natural bipartite graphs, namely the number of top vertices has the same order of magnitude as the number of bottom vertices and there are some high-degree top nodes (see below and Figure 7).

First notice that the decomposition scheme is nothing but a clique covering problem: it computes a set of cliques (which will correspond to the top nodes in the bipartite graph) such that each edge belongs to at least one clique (which ensures that the $\perp$-projection of the decomposition is exactly the original graph). Simple ideas to cover the graph with cliques might be to consider each edge as a clique, or to consider all maximal cliques. However, the first approach would not yield large cliques while the second one could yield too many cliques (the number of maximal cliques may be exponential).

To reach our goal, we proposed [34] the following decomposition. We pick for each edge a largest clique containing it: a clique whose size is maximal among the ones containing the edge. Notice that this clique may contain only two vertices. Moreover, if there are several such cliques for the same edge, we pick one at random. This decomposition ensures the complete covering of the graph. Moreover, the number of cliques is at most equal to the number of edges, which is of the order of the number of vertices. Finally since we take largest cliques, we expect to find most of the large cliques contained in the graph.

In the case of Figure 3 we obtain several cliques of size 2 (namely $\{C, E\}$ and $\{D, F\}$), and we have to choose at random between $\{A, B, C\}$ and $\{B, C, D\}$ when considering the edge $\{B, C\}$. However, these two cliques are obtained from other edges, and we finally obtain a unique decomposition which is nothing but the bipartite graph on the left of the figure.

The central aim of our decomposition scheme is, given the $\perp$-projection of a natural bipartite graph, to produce an artificial bipartite graph similar to the original bipartite graph itself. A way to evaluate it is therefore to decompose the $\perp$-projection version of a natural bipartite complex network and to compare the obtained bipartite network to the original one. This is what we do in Figure 4.

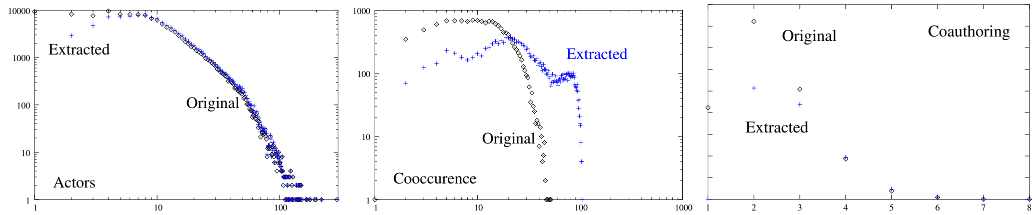

Figure 4: Original clique size distribution for *Actors*, *Cooccurrence* and *Coauthoring*, and extracted clique sizes distribution with the decomposition scheme.

The obtained distributions display some differences for the three graphs decomposed. First, the decomposition scheme produces no cliques of size 1 since the smallest extracted element is the edge (a 2-clique). Moreover, many 2-cliques have not been found, which means that these 2-cliques are not maximal in the original graph. For cliques of size more than 2, our extraction algorithm has been able to find most cliques, or even more. Such new large cliques are induced by the overlapping of other cliques. Notice that in the case of *Cooccurrence* there are many new very large cliques which have been created by overlapping, while in *Actors* and *Coauthoring* this phenomena is very weak. Despite these differences, the obtained size distributions are similar to the original ones. In particular, we obtain a nontrivial number of large cliques and a similar number of cliques, which are the two main points here.

Practical computation

Notice that minimizing the number of cliques leads to the *minimal clique covering problem* which is known to be NP-complete [47, 57]. Computing maximal cliques of a graph is also NP-complete [1, 13] and so is the computation of the largest clique containing a given edge $\{u, v\}$. However, some heuristics make it possible to compute it if the graph is not too large. In our case, we use the following remarks. Let us denote the sets of neighbors of a vertex and an edge by $N(u) = \{v \in V | \{u, v\} \in E\}$ and $N(u, v) = N(u) \cap N(v)$ respectively. First notice that a largest clique containing $\{u, v\}$ in G is also a largest clique containing $\{u, v\}$ in the sub-graph of G induced by $N(u, v) \cup \{u, v\}$. Moreover, if we denote by $\mathcal{C}$ the largest clique in the sub-graph of G induced by $N(u, v)$, then $\mathcal{C} \cup \{u, v\}$ is the clique we are looking for. Figure 5 illustrates this process.

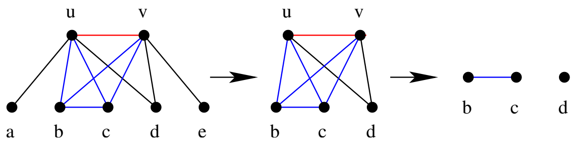

Figure 5: Given a graph $G = (V, E')$, we are looking for a largest clique containing the edge $\{u, v\}$. This clique is necessarily contained in the subgraph induced by $N(u) \cup N(v) \cup \{u, v\} = \{b, c, d, u, v\}$. It is actually sufficient to compute the largest clique $\mathcal{C}$ in the subgraph induced by $N(u) \cap N(v) = N(u, v) = \{b, c, d\}$ since the clique we are looking for is nothing but $\mathcal{C} \cup \{u, v\}$ which, in our case, gives $\{u, v, b, c\}$

Recall that the decomposition process relies on a NP-complete problem in general. However, we observed that in real-world complex networks, the subgraphs induced by $N(u, v)$ for all edges $\{u, v\}$ are in general very dense and very small (Figure 6), which is due to the high clustering and to the power law degree distribution, respectively. This makes it possible to compute the largest clique containing $\{u, v\}$ very efficiently in practice.

Properties of the bipartite graphs

Given the general decomposition scheme, we can now transform any complex network into a bipartite graph. Figure 7 shows the top and bottom degree distribution for the natural bipartite networks *Actors*, *Cooccurrence* and *Coauthoring*, and the ones obtained for *Internet*, *Web* and *Proteins* graphs using our decomposition scheme.

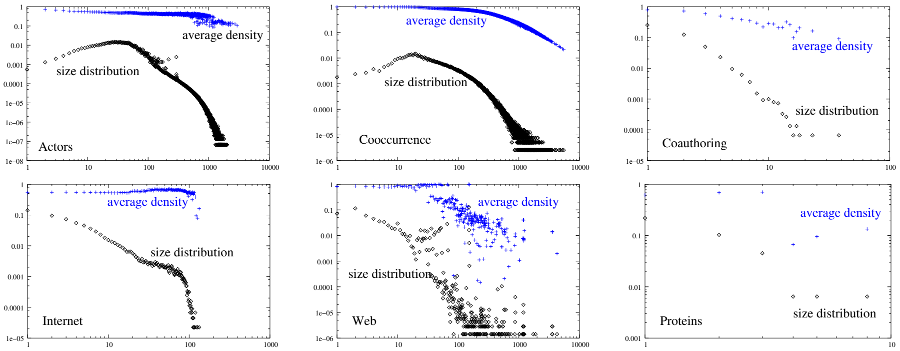

Figure 6: Distribution of the $N(u, v)$ sizes for all edges (u, v), and average density of neighborhoods of given size.

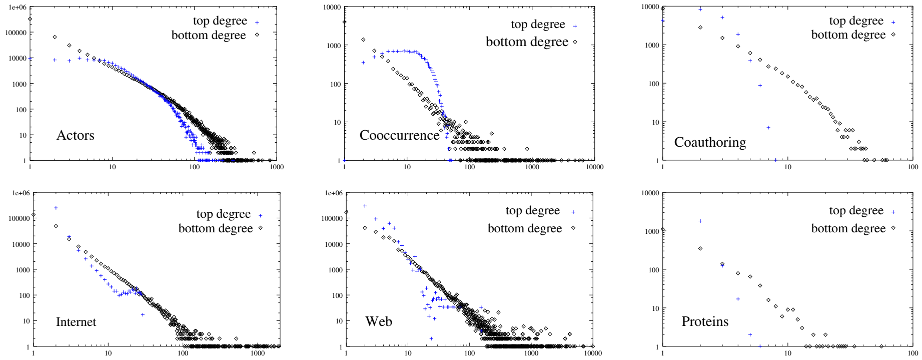

Figure 7: Top and bottom degree distributions for the natural bipartite versions of *Actors*, *Cooccurrence*, and *Coauthoring*, and for the bipartite version of *Internet*, *Web*, and *Proteins* obtained with the decomposition scheme.

All these distributions have a property in common: bottom degree distributions fit very well power laws in all cases. On the contrary, the top degree distributions are of two kinds: while *Cooccurrence*, *Coauthoring*, *Internet* and *Proteins* ones exhibit a Poisson behavior, *Actors* and *Web* ones are more heavy tailed.

These results lead to several remarks. First, the power law bottom degree distribution seems universal, just like the power law distribution in the classical versions of these graphs. Second, the top degree distributions can be qualitatively different and this point is important in the use of the bipartite structure for modeling complex networks since it can impact on some characteristics of the generated graphs. Further remarks will be pointed out in Section 6.

One may also wonder if the degree of a bottom vertex in the bipartite graph and the classical version of the same graph are related. Notice first that the degree of a vertex in the classical graph is the sum of the degrees of the top vertices to which it is connected in the bipartite graph, minus the number of vertices in common in the neighborhood of these vertices. One can easily be convinced that this overlap between neighborhoods, if any, can have a great impact on the degree distribution. To deepen this notion of overlap, one can observe the correlation between the bottom vertex degrees in both bipartite and classical version of the same graph (Figure 8). There exists nontrivial correlations in both cases, which are particularily strong in the case of *Cooccurrence*. Others remarks on the overlap will be discussed further in Section 6.

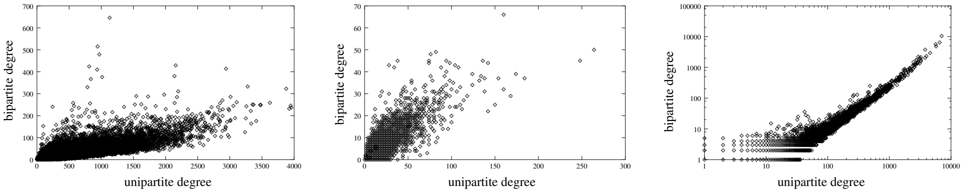

Figure 8: Correlations between degrees in the classical network and the bipartite one. From left to right: *Actors*, *Coauthoring*, and *Cooccurrence*.

Finally, we have shown in this section that *all* complex networks have a nontrivial underlying bipartite structure, which can be computed using our decomposition scheme. This leads us to the following question: is it possible to see the main properties of real-world complex networks as consequences of their underlying bipartite structure? We answer this question in the next sections.

3 The bipartite models.

Our aim is now to use the new general property of real-world complex networks discovered in the previous section, namely their underlying bipartite structure, as a way to propose a model which captures the main wanted properties.

As discussed in the first section of this paper, there are basically two ways to achieve this goal. First, we may try to sample random bipartite graphs with prescribed (top and bottom) degree distributions. Second, we may try to propose a construction process similar to the ones observed in practice, to obtain a *growing* model.

We proposed such models in [34]. In order to deepen the understanding of these models, we here recall and discuss more precisely their definitions, and we provide a full (both analytic and experimental) analysis in the next sections.

Random sampling of bipartite graphs with prescribed degree distribution

One can sample uniformly a random bipartite graph with prescribed (top and bottom) degree distributions as follows (see Figure 9) [17, 55, 56]:

1. generate both top and bottom vertices and assign to each vertex a degree drawn from the given distributions,

2. create for each vertex as many connection points as its degree,

3. link top and bottom connection points randomly,

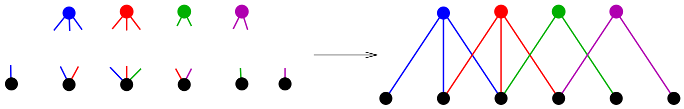

Figure 9: Construction of a random bipartite graph with prescribed degree distributions: first top and bottom vertices are drawn and each vertex is assigned a degree with respect to the given distributions, then edges are chosen randomly between the two sets.

This process generates random bipartite graphs uniformly within the set of bipartite graphs with the given degree distributions. However it cannot be used without taking care of the following constraints: top and bottom distributions cannot be arbitrary since they must allow the total degree of both sets to be equal. Actually one only has to ensure that the number of vertices times the mean of the distribution brings the same value for top and bottom sets. The second problem arises with the fact that even if the two distributions are theoretically consistent, two sets of degrees experimentally drawn from these distributions can be inconsistent (the sums of the degrees are different). A classical trick, which induces no bias, consists in dropping one top and one bottom vertex at random and redraw their degree [55, 56]. This last step may have to be done more than once before one obtains correct values but finally the implementation and use of the model is very simple and efficient [17].

Note that, just like with the MR model, multiple edges may appear. Again, one can easily show that they can be neglected when the graph is large. Moreover, some approaches exist [38, 44] which can easily be modified to obtain random bipartite graphs without multiples edges. This is however out of the scope of this paper.

Growing bipartite model with preferential attachment

The random bipartite model assumes that two distributions, for both top and bottom degrees, are explicitly given. One can also use other rules (preferential attachment for instance) to define them implicitly and introduce a growing model. Indeed, as already noticed, the bottom degree distributions follow a power law. This leads to the following model: at each step, a new top vertex is added and its degree d is sampled from a prescribed (top) distribution (which qualitatively varies between graphs). Then, for each of the d edges of the new vertex, either a new bottom vertex is added (with probability $1-\lambda$) or one is picked among the preexisting ones using preferential attachment (with probability λ). The parameter λ is the *overlap ratio*, defined as the average ratio of preexisting bottom vertices to which a new top vertex is connected.

It is generally not possible to know exactly the order in which cliques are created on real-world bipartite graphs, but the average ratio can be computed globally as $\lambda = 1 - \frac{|\bot|}{\sum d_\top}$. One can compute it and get 0.733 for *Actors*, 0.877 for *Coauthoring* and 0.949 for *Cooccurrence*. Notice that $1-\lambda$ can be rewritten and is simply the inverse average bottom degree (since $\sum d_\top = \sum d_\bot$), therefore a high overlap ratio yields a high average bottom degree (since only few nodes are created at each time step).

At each step of the construction process, the bipartite graph has the required degree distributions: the prescribed top degree distribution is obtained by construction while the power law degree distribution is obtained using preferential attachment, which can be shown formally in exactly the same way as in the original AB model [3]. Notice moreover that this construction process is very similar to the one observed in some real-world cases. For instance, *Actors* is built exactly this way: when a new movie is produced (which corresponds to the addition of a top vertex), it is linked to actors according to their popularity, and to some new actors, playing in a movie for the first time.

Bipartite models and classical graphs

Both models can be defined in the classical (in opposition to *bipartite*) framework in a very similar way (we consider here a graph which can be viewed as the $\perp$-projection of an underlying bipartite graph). Starting with n disconnected vertices, one then iterates the following operation: add all the edges between k vertices, where k is drawn from a given distribution (corresponding to the top degree distribution of the underlying bipartite graph) and where the vertices are chosen with respect to a specific rule (uniformly or using preferential attachment according to their current degree, for instance). See Figure 10. When the vertices are chosen uniformly at random, this model is equivalent to the bipartite one where the bottom distribution is a Poisson law. Notice that if k is always taken equal to 2, then only single edges are added and so if the vertices are chosen uniformly at random we obtain the classical random graph model [29, 12].

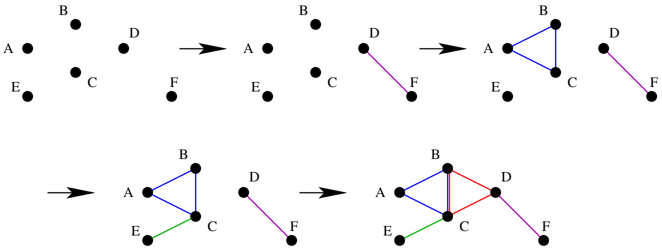

Figure 10: Unipartite version of the model: nodes are initially disconnected and at each step a clique is added on a randomly chosen set of bottom nodes.

A growing model can also be defined in which new vertices are created and cliques are added by choosing a certain amount of pre-existing vertices and some new ones (the cast of a new movie contains some known actors and some new ones). This growing model can also include the preferential attachment rule to choose old vertices. One then obtains the AB model as a special case.

We finally have two models to produce bipartite networks similar to the ones obtained from real-world complex networks, in terms of top and bottom degree distributions. The next question is to ask if they capture the other properties of interest in their $\perp$-projection, namely the average distance, the degree distribution and the clustering. We will answer positively to this question with formal arguments and with experimental results in the next sections.

4 Analysis of the models.

Our aim in this section is to give formal proofs for the main properties of the $\bot$-projection of a random bipartite graph with prescribed degree distributions. Some of these properties, and others, have been studied independently in [55, 56] with different techniques and a different point of view. We however believe that our proofs give new insight on these properties, therefore we give them below. In particular, our proof techniques may be considered as more mathematically rigorous.

Since these properties are induced by a *typical* graph (this is what random sampling gives us), this is a way to answer the following question: what properties are induced by the underlying bipartite structure? In particular, can we see the main properties of real-world complex networks, namely low average distance, power law degree distribution and high clustering, as consequences of the underlying bipartite structure?

We will see that it is indeed the case. Notice that many other properties, like the size distribution of the connected components for instance, are of high interest. It is shown in [55] that under reasonable conditions on the degree distributions the $\bot$-projection is connected, or at least has a giant component. In all the practical cases, these conditions are fulfilled, therefore we will restrict ourselves to this case.

Degree distribution

Let us first consider the degree distribution of the $\bot$-projection of a random bipartite graph $G = (\top, \bot, E)$. Given a bottom vertex u, we denote by $d(u)$ the degree of u in the bipartite graph, and by $d_U(u)$ its degree in its $\bot$-projection. We want to study the distribution of $d_U(u)$ (we actually deal here with the expected value for a randomly chosen u).

Lemma 1 *Let us consider a bottom vertex $u \in \bot$. The expected number of bottom vertices which have a neighbor (in $\top$) in common with u,* i.e. $d_U(u)$, *is:*

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

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

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

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

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

Therefore:

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

which is the formula of the claim. □

This lemma makes it possible to compute the probability for a vertex u in the $\perp$-projection graph 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 the bottom degree distribution follows a power law, the degree distribution in the $\perp$-projection of the graph also follows a power law with the same exponent, which is indeed the case in practice as one can check in Figures 11 and 7.

Average distance

To study the average distance in the $\bot$-projection of a graph obtained with the model, we will use a result from L. Lu about the diameter (*i.e.* the largest distance between any two vertices) of some specific random graphs:

Theorem 1 *[39] Let $G = (V, E)$ be a graph whose vertices are weighted with weights $w_1, \cdots, w_n$, such that each edge $\{i, j\}$ appears with probability $w_i \cdot w_j \cdot p$. If the degrees of the vertices in V follow a power law with an exponent β strictly greater than 2, then the diameter of the graph G is almost surely* $\Theta(\log(n))$[6].

This theorem, together with the one presented above on the degree distribution of the $\bot$-projection of the graph, leads to the following result:

Theorem 2 *Let $G = (\top, \bot, E)$ be a bipartite graph such that the bottom degree distribution follows a power law with an exponent greater than 2. Then the diameter of the $\bot$-projection of G is almost surely* $\Theta(\log(|\bot|))$.

Proof. Given two bottom vertices u and v in $\bot$, the probability that they are connected in the $\bot$-projection is equal to the probability that they are both linked to a same top vertex in G. This probability is exactly proportional to $d_\bot(u) \cdot d_\bot(v)$. Therefore we can apply Theorem 1 considering that the weight of each vertex is its degree and so the connection probability is ensured, and as long as bottom degree distribution follows a power law with an exponent β strictly greater than 2. The diameter of the $\bot$-projection of the graph therefore is almost surely $\Theta(\log(|\bot|))$. □

Since the diameter is an upper bound for the average distance, this theorem implies that the average distance of the $\bot$-projection scales at most as fast as the logarithm of its number of nodes. Notice that, as in the case of random networks [12, 18, 21, 28, 39, 55, 56], the average distance may grow even slower.

Clustering

Recall that the clustering of a vertex v of degree at least 2 in a graph is the probability that two of its neighbors are linked [64], *i.e.* the number of triangles to which v belongs over the number of connected triples centered on it: $c(v) = \frac{|\triangle(v)|}{|\wedge(v)|}$. Then the clustering of the graph is defined as: $\frac{1}{N} \sum_{v, d(v)>1} c(v)$.

[6] One denotes by $f = \Theta(g)$ the fact that $f = \mathcal{O}(g)$ *and* $g = \mathcal{O}(f)$i.

We define the clustering of a vertex restricted to a part of its neighborhood as its clustering in the subgraph induced by this part of its neighborhood.

Hereafter we give a lower bound for the clustering of a graph G' which is the $\bot$-projection of a bipartite graph $G = (\top, \bot, E)$ obtained using the random bipartite model. We show that, under reasonable assumptions on the top and bottom degree distributions, it is bounded by a value independent of the size of the graph. This shows that the model produces graphs with nontrivial clustering.

Before entering in the core of this section, notice that an approximation formula for the clustering of such a graph is given in [55, 56]. Here we give an exact formula for a *lower bound*. Both are interesting since the first one gives an expected value which is indeed very close to the real value, while the second one gives a guaranty that the exact value is above the given quantity. We used this approach because we seek qualitative results only, and so it is sufficient for us to show that the clustering does not tend to 0 when the size of the graph grows. The lower bound achieves this goal.

First notice that the probability for two top vertices to have more than just one bottom vertex in common in their neighborhood tends to zero when the size of the graph grows. We therefore consider any vertex b in the $\bot$-projection of the graph and we suppose that its neighborhood is composed of a set of disjoint cliques. We will prove the following:

- the effect of the number of top vertices of degree 2 to which b is connected on its clustering is negligible, and
- the clustering of b can be bounded by a value which depends only on its degree.

Lemma 2 *Let $\top_{>2}$ denote the set of top neighbors of b in G with degree strictly greater than 2, and $\bot_{>2}$ denote the set of bottom neighbors of $\top_{>2}$. Let p be the fraction of neighbors of b which belong to $\bot_{>2}$, and α be the clustering of b (in G') restricted to $\bot_{>2}$.*

Then the clustering of b in G' scales as $p^2 \cdot \alpha$.

Proof. The fact that the clustering of b restricted to $\bot_{>2}$ is α implies that $|\triangle_{\bot_{>2}}(b)| = \alpha \cdot \binom{p \cdot d}{2}$. If we consider the whole neighborhood of b, instead of just $\bot_{>2}$, the number of triangles does not change while the number of connected triples increases:

$$\begin{aligned} c(b) &= \frac{\alpha \cdot \binom{p \cdot d}{2}}{\binom{d}{2}} \\ &= \alpha \cdot \frac{p \cdot d((p \cdot d - 1)}{d(d-1)} \\ &\sim p^2 \cdot \alpha \end{aligned}$$

which is the formula of the claim. □

Therefore, as long as p is a constant, one can neglect the top vertices of degree 2 when computing the clustering of a given vertex. Now let us prove that the clustering of a bottom vertex can be related to its degree.

Lemma 3 *If b is connected only to top vertices of degree at least 3 in G, then:*

$$c(b) \geq \frac{1}{2 \cdot d(b) - 1}$$

Proof. Suppose b is connected to two top vertices, t_1 and t_2, of degree at least 3 (we deal with the general case below). Then the clustering of b is:

$$c(b) = \frac{\binom{d(t_1)-1}{2} + \binom{d(t_2)-1}{2}}{\binom{d(t_1)+d(t_2)-2}{2}}$$

Suppose now that b is connected to t_2 and t_1' such that $d(t_1') = d(t_1) + 1$, then the clustering of b is:

$$c'(b) = \frac{\binom{d(t_1)+1-1}{2} + \binom{d(t_2)-1}{2}}{\binom{d(t_1)+d(t_2)-1}{2}}$$

and:

$$\begin{aligned} c'(b) - c(b) &= \frac{2 \cdot (d(t_2) - 1)}{(d(t_1) + d(t_2) - 2) \cdot (d(t_1) + d(t_2) - 3)} \\ &> 0 \end{aligned}$$

which means that the clustering grows with the degree of t_1 and t_2. A lower bound for the clustering of b can therefore be obtained when both t_1 and t_2 have the smallest possible degree, 3.

This can be extended to the case where b has more than two top neighbors to obtain the following lower bound:

$$
\begin{aligned}
c(b) &= \frac{\sum_{t_i} \binom{d(t_i)-1}{2}}{\binom{\sum_{t_i}(d(t_i)-1)}{2}} \\
&\geq \frac{\sum_{t_i} \binom{3-1}{2}}{\binom{\sum_{t_i}(3-1)}{2}} \geq \frac{1}{2 \cdot d(b) - 1}
\end{aligned}
$$

which is the formula of the claim. □

The clustering of the classical graph G' can now be easily approximated:

$$c(G') \sim \frac{1}{N} \sum_{b \in \perp} \frac{1}{2d(b) - 1}$$

As long as there is a linear number $c \cdot N$ of vertices b of degree 2, the sum scales linearly with N: $\sum_{b \in \perp} \frac{1}{2 \cdot d(b) - 1} \geq \sum_{b, d(b)=2} \left(\frac{1}{2 \cdot 2 - 1}\right) = \frac{c \cdot N}{3}$ (we could have considered vertices of any constant degree k instead of 2). Therefore the lower bound for the clustering is independent of N. This holds in particular for power law networks since the number of vertices of degree 2 is of the order of $N \cdot 2^{-\alpha}$.

Since we do not consider top vertices of degree 2 in the last formula (due to Lemma 2), we must also ensure that the number of such top neighbors represent at most a constant fraction (not tending to 1) of the neighbors. This is indeed the case for most distributions and in particular for the ones met in practice. We finally obtain that the clustering of the graph is larger than a non-zero constant independently of the size of the graph, which was our aim.

5 Experimental results

The formal results of the previous section give a precise intuition on how the random bipartite graph model with prescribed degree distributions behaves. We can also check its properties experimentally by generating graphs using this model and the same parameters as the ones measured on real-world complex networks. This is what we do in this section with our six examples, for the purely random bipartite model as well as for the one with preferential attachment.

Table 4 and 5 give the values obtained for the average distance and the clustering. Figure 11 shows a comparison between the degree distributions of the original graphs, and the ones obtained with the two bipartite models.

	Internet	Web	Actors	Co-auth	Co-occur	Protein
d	5.80	7	3.6	7.18	2.13	6.74
d_{ER}	5.25	5.47	2.97	7.57	2.06	10.4
d_{MR}	3.25	4.48	2.95	5.77	2.36	5.73
d_{AB}	4.15	5.1	2.93	5.5	2.38	8.15
d_{WS}	5.90	11.23	2559 (*)	2269 (*)	55.6 (*)	509 (*)
d_{rb}	2.97	3.2	3.06	5.07	2.06	5.8
d_{gb}	2.81	3.53	2.83	3.98	2.6	5.45

Table 4: Average distance of the commonly used models and the bipartite models. For each network, we give its actual average distance, and the one obtained with the purely random model d_{ER}, the random model with prescribed degree distribution d_{MR}, the AB model d_{AB}, the WS model d_{WS}, the random bipartite model with prescribed degree distributions d_{rb}, and the growing one with preferential attachment d_{gb}. In the cases pointed by a star (*), the distance is not relevant due to the high clustering.

	Internet	Web	Actors	Co-auth	Co-occur	Protein
c	0.171	0.466	0.785	0.638	0.822	0.153
c_{ER}	0.0001	0.00002	0.0002	0.0002	0.009	0.001
c_{MR}	0.0694	0.017	0.0057	0.001	0.26	0.007
c_{AB}	0.0024	0.0005	0.0015	0.003	0.028	0
c_{WS}	0.171	0.461	0.74 (*)	0.523 (*)	0.74 (*)	0.06 (*)
c_{rb}	0.32	0.663	0.767	0.542	0.831	0.187
c_{gb}	0.65	0.708	0.793	0.632	0.768	0.244

Table 5: Clustering obtained with the commonly used models and the bipartite models. For each network, we give its actual clustering and the clustering obtained with the ER model c_{ER}, the MR model c_{MR}, the AB model c_{AB}, the WS model c_{WS}, and also the random bipartite model with prescribed degree distributions c_{rb}, and the growing one with preferential attachment c_{gb}. Recall that in the cases pointed by a star (*), the real clustering is too large to be obtained with the WS model. Therefore we used in these cases the parameters inducing the maximal clustering.

As expected from the previous section, the graphs we obtain with the random bipartite model have a power law distribution of degrees, a small average distance and a high clustering. Moreover, by definition, they have the same distribution of cliques size as the original network. Therefore the model is qualitatively accurate for the modeling of general real-world complex

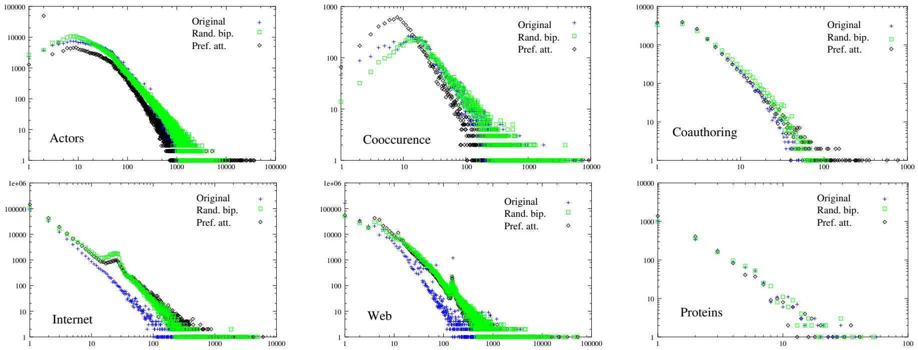

Figure 11: The original degree distribution of our six examples, together with the ones obtained with the random bipartite model and with the growing bipartite model.

networks: the simulations fit real-world values qualitatively well for both clustering and average distance, which proves the relevance of the underlying bipartite structure as an essential property to characterize real-world complex networks.

There are however differences between the values obtained from the bipartite models and real-world networks. They are consequences of the following fact: in the original bipartite networks (both natural ones and the ones obtained from the decomposition), many top nodes have a large neighborhood intersection. In other words, the overlap between cliques is large (if two cliques have one neighbor in common, they certainly have many). This behavior can be viewed as a *bipartite clustering* and is not captured by the bipartite models. The random linking implies that most cliques have only one vertex in common, if any. This is responsible for both the inaccuracy of the models concerning some clusterings and for the irregularities one can observe on some distributions. Figure 12 plots the distribution of the overlap between cliques. This overlap is very small for all random bipartite graphs while it is non trivial for the original graphs.

More precisely, in the case of *Internet*, we noticed the presence of a sub-graph of only 94 vertices which contains all the 494 cliques of size 14 and more. This makes this sub-graph very dense, which implies that the clustering of each of the 94 nodes is very high. However, they have almost no impact on the clustering of the whole graph (due to the average). On the other hand, in the $\perp$-projection of random bipartite networks, these large cliques are disseminated all over the graph which brings two artifacts: there are a lot of vertices having a degree between 14 and 29 which explains the bump on degree distribution (a similar phenomenon can be observed on *Web*), and

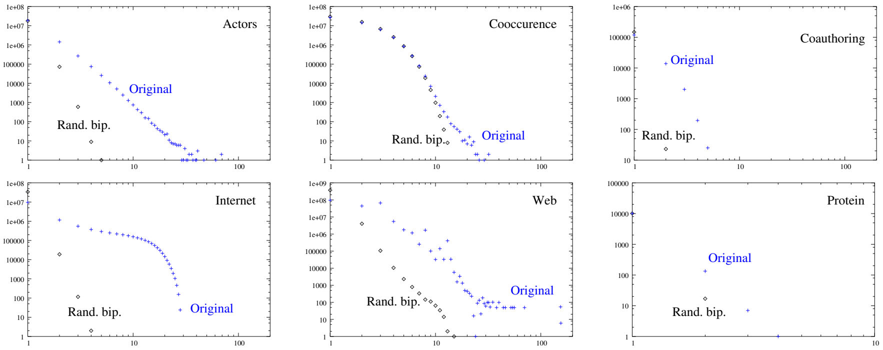

Figure 12: Distribution of the size of the overlap between cliques (*i.e.* intersections of neighbors of top nodes) for both original networks (or decomposed ones) and random bipartite ones.

the number of vertices with high clustering is drastically increased because of their presence in large cliques (from 94 to $50,000$).

These experimental results should also be compared to the ones obtained with the currently most used models, presented in Section 1. This comparison gives evidence for the fact that the models we propose may be considered as an important step towards the realistic modeling of complex networks.

All these remarks hold both for the growing bipartite model and for the random one. This is worth to notice, since it may be very important in some contexts that the model produces *growing* graphs with realistic properties, and in other contexts that the obtained graphs are representative of a precise class of graphs.

6 Conclusion and discussion

In this paper, we propose bipartite graphs as a general tool for the modeling of real-world complex networks. They make it possible to achieve the following challenges:

- the obtained networks have the three main wanted properties (logarithmic average distance, high clustering and power law degree distribution),
- the models are based on a *realistic* construction process representative of what happens in some real-world cases, and

- their definitions are simple enough to make it possible to give some intuition and some proofs of their properties.

Moreover, they can be derived in two versions: one which relies on random sampling among a class of graphs, and one which relies on an iterative construction process. This makes them suitable for a wide variety of usages.

Whereas many models have already been introduced, this one is the first which reaches all these goals at the same time. In this sense, it may be considered as a significant step towards the realistic modeling of complex networks. Moreover, it is very simple to obtain graphs using this model (we provide a generator at [17]), which makes it highly suitable for simulation purposes.

The model is based on the discovery that all real-world complex networks have an underlying bipartite structure which can be seen as responsible for their main properties. Some networks naturally have this structure. For the others, we show that they 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 general behavior of complex systems.

However, as already stressed in previous sections, the overlapping between cliques is not taken into account by the bipartite model which in some way distributes cliques all over the networks independently of the nodes implied. On the contrary, it seems obvious that graphs such as *Actors* are not randomly constructed: actors from a same country are more likely to play together, for instance. This lack of overlapping can also be described on the bipartite graphs: if two top nodes have more than one bottom node in the intersection of their neighborhood, then this yields a non trivial bipartite clique. On the other hand, for the graphs generated with both bipartite models, most such bipartite cliques are trivial ones (as long as there are no too many cliques).

An analogy can be made with the clustering in random graphs (ER graphs for instance), in which neighborhoods of vertices are very sparse while real-world neighborhoods are quite dense: one could say that real-world bipartite networks are bi-clusterized while random ones are not, even if they capture the most common properties.

There are many directions in which this work may be extended. Solving the previous drawback is one of them. This model might also be extended to the case of directed and weighted graphs. These problems rely on giving a new definition to the concept of clique which can be used in this context.

Another similar problem occurs when the graph is only partially known. In this case, some edges are missing, which might yield to only trivial cliques.

A solution to this problem could be to study a model with quasi-cliques, that is cliques with some missing links/nodes. Embedding this concept in the bipartite vision however is nontrivial and remains to be done.

One may also use this model to deepen the study of some phenomena of high interest like the robustness of networks, the spread of rumors and diseases, etc. The random graph model with prescribed degree distribution already led to important advances on these questions [19, 20, 52, 58]. They should now be extended to the bipartite models in order to evaluate the impact of clustering on these problems. We argue that this is a strength of our approach since results on random graphs with prescribed degrees can be directly adapted to our model in order to take the clustering into account.

Finally, let us emphasize on the fact that the study of real-world complex networks is only at its beginning. The discovery of their statistical properties, the analysis of the impact of these properties, their integration into accurate models, and the use of these models in simulation and analysis are key issues for our understanding of real-world complex networks, which has crucial fundamental and applicative implications. Our work lies in this context. It proposes a solution to the problem of the realistic random modeling of real-world complex networks (in the sense of the three main observed properties), and it points out some relevant directions for further research.

Acknowledgments. We thank Annick Lesne, Clémence Magnien and James Martin for careful reading of preliminary versions and useful comments. We also thank the anonymous referees for helpful comments.

References

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

[2] W. Aiello, F.R.K. Chung, and L. Lu. A random graph model for massive graphs. In *ACM Symposium on Theory of Computing (STOC)*, pages 171–180, 2000.

[3] R. Albert and A.-L. Barabási. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[4] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics 74, 47*, 2002.

[5] R. Albert, H. Jeong, and A.-L. Barabási. Diameter of the world wide web. *Nature*, 401:130–131, 1999.

[6] R. Albert, H. Jeong, and A.-L. Barabási. Error and attack tolerance in complex networks. *Nature*, 406:378–382, 2000.

[7] arXiv.org e Print archive. http://arxiv.org/.

[8] A.-L. Barabási, E. Ravasz, and T. Vicsek. Deterministic scale-free networks. *Physica A 299, (3-4)*, pages 559–564, 2001.

[9] E.A. Bender and E.R. Canfield. The asymptotic number of labeled graphs with given degree sequences. *J. Combin. Theory Ser. A*, 24:296–307, 1978.

[10] M. Boguna, R. Pastor-Satorras, and A. Vespignani. Epidemic spreading in complex networks with degree correlations. In al J.M. Rubi et, editor, *XVIII Sitges Conference "Statistical Mechanics of Complex Networks"*. Springer Verlag, 2003.

[11] B. Bollobás. A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. *Europ. J Combinatorics*, 1:311–316, 1980.

[12] B. Bollobás. *Random Graphs*. Academic Press, 1985.

[13] I. Bomze, M. Budinich, P. Pardalos, and M. Pelillo. The maximum clique problem. In D.-Z. Du and P. M. Pardalos, editors, *Handbook of Combinatorial Optimization*, volume 4. Kluwer Academic Publishers, Boston, MA, 1999.

[14] A.Z. Broder, S.R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. L. Wiener. Graph structure in the web. *WWW9 / Computer Networks*, 33(1-6):309–320, 2000.

[15] D.S. Callaway, M.E.J. Newman, S.H. Strogatz, and D.J. Watts. Network robustness and fragility: Percolation on random graphs. *Phys. Rev. Lett.*, 85:5468–5471, 2000.

[16] 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.

[17] Source code for the random bipartite graph generator. http://www.liafa.jussieu.fr/~guillaume/programs/.

[18] R. Cohen, D. ben Avraham, and S. Havlin. *Handbook of graphs and networks*, chapter 4: Structural properties of scale free networks. Wiley-VCH, 2002.

[19] R. Cohen, K. Erez, D. ben Avraham, and S. Havlin. Resilience of the internet to random breakdown. *Phys. Rev. Lett.*, 85:4626–4628, 2000.

[20] R. Cohen, K. Erez, D. ben Avraham, and S. Havlin. Breakdown of the internet under intentional attack. *Phys. Rev. Lett.*, 86:3682–3685, 2001.

[21] R. Cohen and S. Havlin. Scale free networks are ultrasmall. *Phys. Rev. Lett.*, 90, 2003.

[22] 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*, pages 73–87.

[23] Self-Organized Networks Database. http://www.nd.edu/~networks/database/index.html.

[24] The Internet Movie Database. http://www.imdb.com/.

[25] S.N. Dorogovtsev and J.F.F. Mendes. Exactly solvable small-world network. *Euro. phys. Lett.*, 50 (1):1–7, 2000.

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

[27] S.N. Dorogovtsev, J.F.F. Mendes, and A.N. Samukhin. Structure of growing networks with preferential linking. *Phys. Rev. Lett. 85*, pages 4633–4636, 2000.

[28] S.N. Dorogovtsev, J.F.F. Mendes, and A.N. Samukhin. Metric structure of random networks. *Nucl. Phys. B 653, 307*, 2003.

[29] P. Erdös and A. Rényi. On random graphs I. *Publ. Math. Debrecen*, 6:290–297, 1959.

[30] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *SIGCOMM*, pages 251–262, 1999.

[31] R. Ferrer and R.V. Solé. The small-world of human language. In *Proceedings of the Royal Society of London*, volume B268, pages 2261–2265, 2001.

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

[33] R. Govindan and H. Tangmunarunkit. Heuristics for internet map discovery. In *IEEE INFOCOM 2000*, pages 1371–1380, Tel Aviv, Israel, March 2000. IEEE.

[34] Jean-Loup Guillaume and Matthieu Latapy. Bipartite structure of *all* complex networks. *Information Processing Letters (IPL)*, 90(5):215–221, 2004.

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

[36] B.J. Kim, C.N. Yoon, S.K. Han, and H. Jeong. Path finding strategies in scale-free networks. *Phys. Rev. E 65, 027103.*, 2002.

[37] J.M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A.S. Tomkins. The Web as a graph: Measurements, models, and methods. In T. Asano, H. Imai, D. T. Lee, S. Nakano, and T. Tokuyama, editors, *Proc. 5th Annual Int. Conf. Computing and Combinatorics, COCOON*, number 1627. Springer-Verlag, 1999.

[38] M. Latapy and F. Viger. Random generation of large connected simple graphs with prescribed degree distribution. In *proceedings of the 11-th international conference on Computing and Combinatorics COCOON'05*, 2005.

[39] L. Lu. The diameter of random massive graphs. In ACM-SIAM, editor, *12th Ann. Symp. on Discrete Algorithms (SODA)*, pages 912–921, 2001.

[40] T. Luczak. Sparse random graphs with a given degree sequence, in Random Graphs, vol. 2. A.M. Frieze, T. uczak eds. Wiley, New York, 1992. pages. 165-182.

[41] D. Magoni and J.-J. Pansiot. Influence of network topology on protocol simulation. In *ICN'01 - 1st IEEE International Conference on Networking*, volume Lecture Notes in Computer Science, pages 762–770, July 9-13, 2001.

[42] M. Mihail and C. Papadimitriou. the eigenvalue power law, 2002.

[43] S. Milgram. The small world problem. *Psychology today*, 1:61–67, 1967.

[44] R. Milo, N. Kashtan, S. Itzkovitz, M.E.J. Newman, and U. Alon. On the uniform generation of random graphs with prescribed degree sequences, 2003. cond-mat/0312028.

[45] M. Molloy and B. Reed. A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms*, pages 161–179, 1995.

[46] M. Molloy and B. Reed. The size of the giant component of a random graph with a given degree sequence. *Combin. Probab. Comput.*, pages 295–305, 1998.

[47] S.D. Monson, N.J. Pullman, and R. Rees. A survey of clique and biclique coverings and factorizations of (0,1)-matrices. *Bull. Inst. Combin. Appl.*, 14:17–86, 1995.

[48] C. Moore and M.E.J. Newman. Epidemics and percolation in small-worlds networks. *Phys. Rev. E*, 61:5678–5682, 2000.

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

[50] M.E.J. Newman. Scientific collaboration networks: I. Network construction and fundamental results. *Phys. Rev. E*, 64, 2001.

[51] M.E.J. Newman. Scientific collaboration networks: II. Shortest paths, weighted networks, and centrality. *Phys. Rev. E*, 64, 2001.

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

[53] M.E.J. Newman. mixing patterns in networks. *Phy. Rev. E*, 67, 2003. cond-mat/0209450.

[54] M.E.J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[55] M.E.J. Newman, D.J. Watts, and S.H. Strogatz. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E*, 2001.

[56] M.E.J. Newman, D.J. Watts, and S.H. Strogatz. Random graph models of social networks. *Proc. Natl. Acad. Sci. USA*, 99 (Suppl. 1):2566–2572, 2002.

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

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

[59] S.H. Strogatz. Exploring complex networks. *Nature 410*, 2001.

[60] Lakshminarayanan Subramanian, Sharad Agarwal, Jennifer Rexford, and Randy H. Katz. Characterizing the internet hierarchy from multiple vantage points. In *Proc. of IEEE INFOCOM 2002, New York, NY*, Jun 2002.

[61] H. Tangmunarunkit, R. Govindan, S. Jamin, S. Shenker, and W. Willinger. On characterizing network hierarchy. Technical Report 03-782, Computer Science Department, University of Southern California, 2001. submitted.

[62] Bible Today New International Version. http://www.tniv.info/bible/.

[63] T. Walsh. Search in a small world. In *IJCAI*, pages 1172–1177, 1999.

[64] D.J. Watts and S.H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393:440–442, 1998.