Basic Notions for the Analysis of Large Two-mode Networks

Matthieu Latapy,[1] Clémence Magnien [1] and Nathalie Del Vecchio [2]

Abstract

Many large real-world networks actually have a 2-mode nature: their nodes may be separated into two classes, the links being between nodes of different classes only. Despite this, and despite the fact that many ad-hoc tools have been designed for the study of special cases, very few exist to analyse (describe, extract relevant information) such networks in a systematic way. We propose here an extension of the most basic notions used nowadays to analyse large 1-mode networks (the classical case) to the 2-mode case. To achieve this, we introduce a set of simple statistics, which we discuss by comparing their values on a representative set of real-world networks and on their random versions. This makes it possible to evaluate their relevance in capturing properties of interest in 2-mode networks.

Introduction.

A bipartite graph is a triplet $G = (\top, \bot, E)$ where $\top$ is the set of *top* nodes, $\bot$ is the set of *bottom* nodes, and $E \subseteq \top \times \bot$ is the set of links. The difference with *classical* graphs lies in the fact that the nodes are in two disjoint sets, and that the links always are between a node of one set and a node of the other. In other words, there cannot be any link between two nodes in the same set.

Many large real-world networks of interest may be modeled naturally by a bipartite graph. These networks are called *2-mode networks*, or *affiliation networks* when they represent groups and members (*i.e.* each link represents a social actor's affiliation to a group). Let us cite for instance the actors-movies network, where each actor is linked to the movies he/she played in (*e.g.*, Watts & Strogatz, 1998; Newman *et al.*, 2001a), authoring networks, where the authors are linked to the paper they signed (*e.g.*, Newman, 2001a; Newman, 2001b), occurrence networks, where the words occurring in a book are linked to the sentences of the book they appear in (*e.g.*, Ferrer & Solé, 2001), company board networks, where the board members are linked to the companies they lead (*e.g.*, Robins & Alexander, 2004; Conyon & Muldoon, 2004; Battiston & Catanzaro, 2004), and peer-to-peer exchange networks in which peers are linked to the data they provide/search (*e.g.*, Le Fessant *et al.*, 2004; Voulgaris *et al.*, 2004; Guillaume *et al.*, 2005; Guillaume *et al.*, 2004).

[1] LIP6 – CNRS and Université Pierre et Marie Curie (UPMC) – 4, place Jussieu, 75005 Paris, France – Firstname.Lastname@lip6.fr

[2] LARGEPA – Université Paris 2 – 13, avenue Bosquet, 75007 Paris, France – nathdelvecchio@yahoo.com

Although there is nowadays a significant amount of notions and tools to analyse (classical) 1-mode networks, there is still a lack of such results fitting the needs for analysing 2-mode networks. In such cases, one generally has to transform the 2-mode network into a 1-mode one and/or to introduce ad-hoc notions. In the first case, there is an important loss of information, as well as other problems that we detail below (Section 3). In the second case, there is often a lack of rigor and generality, which makes the relevance of the obtained results difficult to evaluate.

The aim of this contribution is to provide a set of simple statistics which will make it possible and easy to analyse real-world 2-mode networks (or at least make the first step towards this goal) while keeping their bipartite nature.

To achieve this, we will first present an overview of the basic notions and methodologies used in the analysis of 1-mode networks. We will then show how people usually transform bipartite networks into 1-mode networks in order to be able to analyse them with the tools designed for this case. This will lead us to a description of the state of the art, then of the methodology used in this paper. Finally, we will present and evaluate the statistics we propose for the analysis of 2-mode networks.

Before entering in the core of this contribution, let us notice that we only deal here with simple [3], undirected, unweighted, static networks. Considering directed, weighted, and/or dynamic networks is out of the scope of this paper; we will discuss this further in the conclusion. Moreover, in all the cases we will consider here (and in most real-world cases), the graph has a huge connected component, *i.e.* there exists a path in the graph from almost any node to any other. In the following, we will make our statistics on the whole graph everywhere this makes sense, but we will restrict ourselves to the largest connected component where this is necessary (namely for distance computations). Again, this is classical in the literature and has no significant impact on our results.

1 Classical notions.

Let us consider a (classical) graph $G = (V, E)$, where V is the set of nodes and $E \subseteq V \times V$ is the set of links. We will denote by $N(v) = \{u \in V,\ (u, v) \in E\}$ the *neighbourhood* of a node v, the elements of $N(v)$ being the *neighbours* of v. The number of nodes in $N(v)$ is the *degree* of v: $d^o(v) = |N(v)|$.

The most basic statistics describing such a graph are its size $n = |V|$, its number of links $m = |E|$, and its average degree $k = \frac{2m}{n}$. Its density $\delta(G) = \frac{2m}{n(n-1)}$, *i.e.* the number of existing links divided by the number of possible links, also is an important notion. It is nothing but the probability that two randomly chosen (distinct) nodes are linked together.

Going further, one may define the distance between two nodes in the graph as the minimal number of links one has to follow to go from one node to the other. Note that this only make sense if there is a path between the two nodes, *i.e.* if they are in the same connected component. As explained above, in all the paper, we will only consider distances between the nodes in the

[3] This means that we do not allow loops (links from a node to itself) nor multiple links between two given nodes. This is classical in studies of large networks: loops are managed separately, if some occur, and multiple links are generally encoded as link weighs, or simply ignored.

largest connected component (and we will give its size). Then, the average distance of the graph, $d(G)$, is nothing but the average of the distances for all pairs of nodes in the largest connected component.

The statistics described above are the ones we will call the *basic* statistics. The next one is not so classical. It is the degree distribution, *i.e.* for all integer i the fraction p_i of nodes of degree i. In other words, it is the probability that a randomly chosen node has degree i. One may also observe the correlations between degrees, defined as the average degree of the neighbours of nodes of degree i, for each integer i. Other notions concerning degrees have been studied, like assortativity (Newman, 2003a) for instance, but we do not detail this here.

The last kind of statistics we will discuss here aims at capturing a notion of overlap: it measures the probability that two nodes are linked together, provided they have a neighbour in common. In other words, it is the probability that any two neighbours of any node are linked together. This may be done using two slightly different notions, both called *clustering coefficient*, among which there often is a confusion in the literature [4]. Both will be useful in the following therefore we discuss them precisely here.

The first one computes the probability, for any given node chosen at random, that two neighbours of this node are linked together. It therefore relies on the notion of clustering coefficient for any node v of degree at least 2, defined by:

$$\mathrm{cc}_\bullet(v) = \frac{|E_{N(v)}|}{\frac{|N(v)|(|N(v)|-1)}{2}} = \frac{2|E_{N(v)}|}{d^o(v)(d^o(v)-1)}$$

where $E_{N(v)} = E \cap (N(v) \times N(v))$ is the set of links between neighbours of v. In other words, $\mathrm{cc}_\bullet(v)$ is the probability that two neighbours of v are linked together. Notice that it is nothing but the density of the neighbourhood of v, and in this sense it captures the local density. The clustering coefficient of the graph itself is the average of this value for all the nodes:

$$\mathrm{cc}_\bullet(G) = \frac{\sum_{v \in V} \mathrm{cc}_\bullet(v)}{|\{v \in V,\ d^o(v) \geq 2\}|}.$$

One may define directly another notion of clustering coefficient of G as a whole as follows:

$$\mathrm{cc}_\vee(G) = \frac{3N_\Delta}{N_\vee}$$

where N_Δ denotes the number of triangles, *i.e.* sets of three nodes with three links in G, and $N_\vee$ denotes the number of connected triples, *i.e.* sets of three nodes with at least two links, in G. This notion of clustering is slightly different from the previous one since it gives the probability, when one chooses two links with one extremity in common, that the two other extremities are linked together.

Both notions have their own drawbacks and advantages. The first one has the advantage of giving a value for each node, which makes it possible to observe the distribution of this value and

[4] Some authors make a difference by calling the first notion *clustering coefficient* and the second one *transitivity ratio*, but we prefer to follow the most classical conventions of large network studies here.

the correlations between this value and the degree, for instance. It however has the drawback of reducing the role of high degree nodes. Moreover, importantly, these definitions capture slightly different notions, which may both be relevant depending on the context. We will therefore use both notions in the following. This is why we introduced two different notations, namely $cc_\bullet$ and $cc_\vee$, which emphasises the fact that one is centered on nodes and the other is centered on pairs of links with one extremity in common.

One may consider many other statistics to describe large networks. Let us cite for instance centrality measures, various decompositions, and notions capturing the ability of each node to spread information in the network. See Wasserman & Faust, 1994; Albert & Barabási, 2002; Newman, 2003b; Bornholdt & Schuster, 2003; Brandes & Erlebach, 2005 for surveys from different perspectives. We will not consider here such statistics. Instead, we will focus on the most simple ones, described above, because they play a central role in recent studies of large networks, which we call post-1998 studies, as we will explain in the next section.

2 One-mode large real-world networks.

Many large real-world networks have been studied in the literature, ranging from technological networks (power grids, internet) to social ones (collaboration networks, economical relations), or from biological ones (protein interactions, brain topology) to linguistic ones (cooccurrence networks, synonymy networks). See Wasserman & Faust, 1994; Albert & Barabási, 2002; Newman, 2003b; Bornholdt & Schuster, 2003; Brandes & Erlebach, 2005 and references therein for detailed examples.

It appeared recently (*e.g.*, Watts & Strogatz, 1998; Albert & Barabási, 2002; Newman, 2003b; Bornholdt & Schuster, 2003) that most of these large real-world networks have several nontrivial properties in common. This was unexpected, and led to an important stream of studies, developping a new kind of network analysis which we will call post-1998 network analysis (as it followed the seminal paper Watts & Strogatz, 1998). This section is devoted to an overview and discussion of these properties (based on the definitions given in previous section), on which the rest of the paper will rely. We will use the same notations as in Section 1.

We are concerned here with large networks only, which means that n is large. In most real-world cases, it appeared that m is of the same order of magnitude as n, *i.e.* the average degree k is small compared to n. Therefore, the density generally is very small: $\delta(G) = \frac{kn}{n(n-1)} \sim \frac{k}{n}$, which is close to 0 since n is much larger than k in general. We will always suppose we are in this case in the following.

It is now a well known fact that the average distance in large real-world networks is in general very small (*small-world* effect), even in very large ones, see for instance Milgram, 1967; Watts & Strogatz, 1998. This is actually true in most graphs, since a small amount of randomness is sufficient to ensure this, see for instance Watts & Strogatz, 1998; Kleinberg, 2000a; Kleinberg, 2000b; Bollobas, 2001; Erdös & Rényi, 1959. This property, though it may have important consequences and should be taken into account, should therefore not be considered as a significant property of a given network (see Section 5).

Another issue which received recently much attention, see for instance Faloutsos *et al.*, 1999;

Barabasi & Albert, 1999, is the fact that the degree distribution [5] of most large real-world networks is highly heterogeneous, often well fitted by a power law: $p_k \sim k^{-\alpha}$ for an exponent α generally between 2 and 3.5. This means that, despite most nodes have a low degree, there exists nodes with a very high degree. This implies in general that the average degree is not a significant property, bringing much less information than the exponent α which is a measurement of the heterogeneity of degrees.

If one samples a random network with the same size (*i.e.* as many nodes and links) as a given real-world one [6], thus with the same density, then the obtained degree distribution is qualitatively different: it follows a Poisson law. This means that the heterogeneous degree distribution is not a trivial property, in the sense that it makes large real-world networks very different from most graphs (of which a random graph is typical). The degree correlations and other properties on degrees, however, behave differently depending on the network under concern.

Going further, the clustering coefficients (according to both definitions) are quite large in most real-world networks: despite most pairs of nodes are not linked together (the density is very low), if two nodes have a neighbour in common then they are linked together with a probability significantly higher than 0 (the local density is high). However, the clustering coefficient distributions, their correlations with degrees, and other properties related to clustering, behave differently depending on the network under concern.

If, as above, one samples a random graph with the same size as an original one then the two definitions of clustering coefficients are equivalent and equal to the density. The clustering coefficients therefore are very low in this case. If one samples a random graph with the same number of nodes *and* the very same degree distribution [7] then the clustering coefficients still are very small, close to 0 (Newman, 2003b). Clustering coefficients therefore capture a property of networks which is not a trivial consequence of their degree distribution.

Finally, it was observed that the vast majority of large real-world networks have a very low density, a small average distance, a highly heterogeneous degree distribution and high clustering coefficients. These two last properties make them very different from random graphs (both purely random and random with prescribed degree distribution). More subtle properties may be studied, but until now no other one appeared to be a general feature of most large real-world networks. The properties described here therefore serve, in most post-1998 studies, as a basis for the analysis of large real-world networks, and so we will focus on them in the following. Our aim will be to define and discuss their equivalent for 2-mode networks / bipartite graphs.

[5] See the appendix, page 30 for more detailed definitions and hints on how to understand this kind of statistics.

[6] We consider here a network chosen uniformly at random among the ones having this size, using typically the Erdös and Rényí model (Bollobas, 2001; Erdös & Rényi, 1959).

[7] We consider here a network chosen uniformly at random among the ones having this number of nodes and this degree distribution, using typically the *configuration* model (Bender & Canfield, 1978; Bollobas, 2001; Molloy & Reed, 1995; Molloy & Reed, 1998; Viger & Latapy, 2005.

3 Projection.

Let us now consider a large 2-mode network modeled as a bipartite graph $G = (\top, \bot, E)$. The $\bot$-projection of G is the graph $G_\bot = (\bot, E_\bot)$ in which two nodes (of $\bot$) are linked together if they have at least one neighbour in common (in $\top$) in G: $E_\bot = \{(u, v),\ \exists x \in \top :\ (u, x) \in E \text{ and } (v, x) \in E\}$. The $\top$-projection $G_\top$ is defined dually. See Figure 1 for an example.

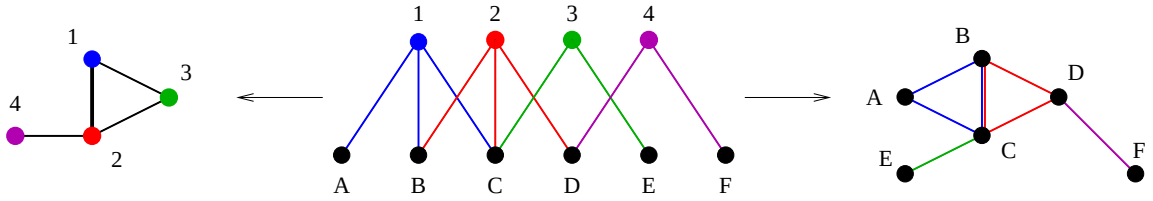

Figure 1: An example of bipartite graph (center), together with its $\top$-projection (left) and its $\bot$-projection (right).

In order to be able to use the many notions defined on 1-mode networks, and to compare a particular network to others, one generally transforms a 2-mode network into its $\bot$-projection, often called the one-mode version of the network. This was typically done for the 2-mode networks we presented in the introduction: the actors-movies network is transformed into its $\bot$-projection where two actors are linked if they acted together in a movie (*e.g.*, Watts & Strogatz, 1998); the authoring networks are transformed into their $\bot$-projections, *i.e.* coauthoring networks where two authors are linked if they signed a paper together (*e.g.*, Newman, 2001a; Newman, 2001b; Newman *et al.*, 2001a); the occurrence networks are transformed into their $\bot$-projections, *i.e.* cooccurrence networks where two words are linked if they appear in the same sentence (*e.g.*, Ferrer & Solé, 2001); the company board networks are transformed into their $\bot$-projections where two persons are linked together if they are member of a same board (*e.g.*, Robins & Alexander, 2004; Conyon & Muldoon, 2004; Battiston & Catanzaro, 2004; Kogut & Walker, 2003; Kogut *et al.*, 2006); and the peer-to-peer exchange networks are transformed into their $\bot$-projections where two data are linked together if they are provided/searched by a same peer (*e.g.*, Le Fessant *et al.*, 2004; Voulgaris *et al.*, 2004; Guillaume *et al.*, 2005; Guillaume *et al.*, 2004).

This approach is of course relevant since the projections under study make sense, and also encode much information. Moreover, this allows the study of 2-mode networks using the powerful tools and notions provided for classical, 1-mode, networks. We however argue that in most cases there would be a significant gain in considering the bipartite version of the data. The main reasons are as follows.

- Most importantly, there is much information in the bipartite structure which may disappear after projection. For instance, the fact that two actors played in many movies together, and the size of these movies, brings much information which is not available in the projection, in which they are simply linked together. This loss of information is particularly clear when one notices that there are many bipartite graphs which lead to the same projection (while each bipartite graph has only one $\top$- and one $\bot$-projection), see Guillaume & Latapy,

2004b; Guillaume & Latapy, 2004a. The fact that much important information is encoded in the bipartite structure is a central point which we will illustrate all along this paper.

- Notice that each top node of degree d induces $\frac{d(d-1)}{2}$ links in the $\perp$-projection, and conversely. This induces an inflation of the number of links when one goes from a bipartite gaph to its projection, see Table 1. In our examples, this is particularly true for peer-to-peer: the number of links reaches more than 10 billions in the $\perp$-projection, which needs more than 80 GigaBytes of central memory to be stored using classical (compact) encodings (while the original 2-mode network needs less than 500 MegaBytes). This is a typical case in which the huge number of links induced by the projection is responsible for limitations on the computations we are able to handle on the graph in practice.

	actors-movies	authoring	occurrences	peer-to-peer
Number of links in G	1,470,418	45,904	183,363	55,829,392
Number of links in $G_\perp$	15,038,083	29,552	392,066	10,142,780,673
Number of links in $G_\top$	20,490,112	134,492	51,405,275	1,085,217,140

Table 1: Number of links in 2-mode networks and their projections, for the four examples we will describe in Section 5.

- Finally, some properties of the projection may be due to the projection process rather than the underlying data itself. For instance, it is shown in Newman *et al.*, 2001a; Guillaume & Latapy, 2004b; Guillaume & Latapy, 2004a that when considering the projection of a random bipartite graph, one observes high clustering coefficients. Therefore, high clustering coefficients in projections may not be viewed as significant properties: they are consequences of the bipartite nature of the underlying 2-mode network. Likewise, the projection may lead to very dense networks, even if the bipartite version is not dense; this is particularly the case here for the $\top$-projection of occurrences.

One way to avoid some of these problems is to use a *weighted* projection. For instance, the weight of a link (u, v) between two bottom nodes in the weighted $\perp$-projection may be defined as the number of (top) neighbours u and v have in common in the bipartite graph. Other definitions may be considered: each top node may contribute to each link it induces in the $\perp$-projection in a way that decreases with its degree, for instance. In all cases, and despite such an approach is relevant and promising, one still loses a significant amount of information, and one transforms the problem of analysing a bipartite structure into the problem of analysing a weighted one, which is not easier. Indeed, despite the fact that important progress has recently be done in this direction (Barrat *et al.*, 2004; Barthélemy *et al.*, 2005; Newman, 2004), much remains to be done before being able to analyse precisely the structure of weighted networks.

Our aim in this paper is to provide an alternative to the projection approach, leading to a better understanding of 2-mode networks. It must however be clear that (weighted) projection approachs also lead to significant insight, and we consider that the two approaches should be used as complementary means to understand in details the properties of 2-mode networks.

4 State of the art.

Two-mode networks have been studied in an amazingly wide variety of context. Let us cite for instance company boards (*e.g.*, Robins & Alexander, 2004; Conyon & Muldoon, 2004; Battiston & Catanzaro, 2004; Newman *et al.*, 2001a), sport teams (*e.g.*, Bonacich, 1972; Onody & de Castro, 2004), movie actors (*e.g.*, Watts & Strogatz, 1998; Newman *et al.*, 2001a), management science (*e.g.*, Kogut & Walker, 2003; Kogut *et al.*, 2006), human sexual relations (*e.g.*, Ergun, 2002; Lind *et al.*, 2005), attendance to events (*e.g.*, Faust *et al.*, 2002; Freeman, 2003), financial networks (*e.g.*, Caldarelli *et al.*, 2004; Dahui *et al.*, 2005; Garlaschelli *et al.*, 2004; Young-Choon, 1998), recommandation networks (*e.g.*, Perugini *et al.*, 2003), theatre performances (*e.g.*, Agneessens *et al.*, 2004; Uzzi & Spiro, 2005), politic ativism (*e.g.*, Boudourides & Botetzagias, 2004), student course registrations (*e.g.*, Holme *et al.*, 2004), word cooccurrences (*e.g.*, Dhillon, 2001; Véronis & Ide, 1995), file sharing (*e.g.*, Iamnitchi *et al.*, 2004; Le Fessant *et al.*, 2004; Voulgaris *et al.*, 2004; Guillaume *et al.*, 2005; Guillaume *et al.*, 2004), and scientific authoring (*e.g.*, Roth & Bourgine, 2005; Morris & Yen, 2005; Newman, 2001a; Newman, 2001b; Newman, 2000).

These studies are made in disciplines as various as social sciences, computer science, linguistics and physics, which makes the literature very rich. In all these contexts, scientists face 2-mode networks which they try to analyse, with various motivations and tools. They all have one feature in common: they insist on the fact that the bipartite nature of their data plays an important role, and should be taken into account. They also emphasise the lack of notions and tools for doing so.

Because of this lack of relevant notions and tools, most authors have no choice but to consider the most relevant projection of their 2-mode network. This leads for instance to studies of interlocks between companies, see Robins & Alexander, 2004; Conyon & Muldoon, 2004, studies of coauthoring networks, see Newman, 2001a; Newman, 2001b; Newman, 2000, or studies of exchanges between peers in peer-to-peer systems, see Le Fessant *et al.*, 2004; Voulgaris *et al.*, 2004; Guillaume *et al.*, 2005; Guillaume *et al.*, 2004.

Many authors realise that this approach is not sufficient, and try to use the bipartite nature of their data. This is generally done by combining the use of projections and the use of basic bipartite statistics, mostly degrees. For instance, one studies the coauthoring relations (typically a projection) and the distributions of the number of papers signed by authors and of the number of authors of papers (*i.e.* the bipartite degree distributions, see Section 6) (Newman, 2000). Authors may also consider weighted projections, see for instance Battiston & Catanzaro, 2004; Morris & Yen, 2005; Guillaume *et al.*, 2004; Guillaume *et al.*, 2005; Iamnitchi *et al.*, 2004; Newman, 2000, which has advantages and drawbacks, as discussed in Section 3.

Going further, some authors introduce bipartite notions designed for the case under study. This is often implicit and restricted to very basic properties, like the case of degree distributions cited above (which essentially capture the size of *events*, and the number of events in which *persons* or *objects* are involved, in most cases). But some authors introduce more subtle notions, like notions of overlap (Bonacich, 1972), clustering (Borgatti & Everett, 1997; Robins & Alexander, 2004; Lind *et al.*, 2005), centrality measures (Faust, 1997), degree correlations (Peltomaki & Alava, 2005), and others (Young-Choon, 1998; Ergun, 2002; Caldarelli *et al.*, 2004; Perugini

et al., 2003; Iamnitchi *et al.*, 2004; Borgatti & Everett, 1997; Robins & Alexander, 2004; Lind *et al.*, 2005). Most of these notions are ad hoc and specific to the case under study, but some of them actually are very general or may be generalised. One of our central aims here is to give a complete and unified framework for the most general of these notions. We will cite appropriate references when the notions we will discuss have already been considered previously.

As already said, a different and interesting approach is developed in Newman *et al.*, 2001a; Guillaume & Latapy, 2004b; Guillaume & Latapy, 2004a. The authors study the expected properties of the projections given the properties (namely the degree distributions) of the underlying bipartite graph. They show in particular that the expected clustering coefficient in the projections is large, and give an efficient estimation formula; this means that a high clustering coefficient in a projection may be seen as a consequence of the underlying bipartite structure rather than a specific property of the network. Conversely, if the clustering coefficient of the projection is different from the expected one, it means that the underlying bipartite structure has nontrivial properties responsible for it. These properties should therefore be further analysed. Our aim here is to propose notions and tools for such an analysis. This approach has been used with profit in several cases, see for instance Newman *et al.*, 2001a; Newman *et al.*, 2002; Conyon & Muldoon, 2004; Uzzi & Spiro, 2005.

Finally, a significant effort has already been made to achieve the goal we have here, or similar goals: some studies propose general approaches for the analysis of 2-mode networks. This is for instance the case of Faust, 1997, focused on centrality measures, of Breiger, 1974, which proposes to consider both projections and compare them, and of Bonacich, 1972, which studies in depth the notion of overlap.

Let us cite in particular Borgatti & Everett, 1997, which has the very same aim as we have here, but belongs to what we call *classical*, or pre-1998, social network analysis. In particular, they do not use the comparison with random graphs, central to our contribution (see Section 5), which probably reflects the fact that this method was not as usual in 1997 as it is now. For the same reasons, they do not deal with clustering questions, which play a key role here. On the other hand, they address some important issues (like visualisation) which we consider as out of the scope of our contribution. It is interesting to see that, although the initially claimed aim is very similar, the final contributions are significantly different.

Other researchers propose formalisms suited for the analysis of 2-mode networks, often based on a generalisation of well known models. Let us cite Galois lattices (*e.g.*, Roth & Bourgine, 2005), correspondence analysis (*e.g.*, Jr., 2000; Faust, 2005), extensions of blockmodels (*e.g.*, Borgatti & Everett, 1992; Doreian *et al.*, 2004) and p* models (*e.g.*, Skvoretz & Faust, 1999; Faust *et al.*, 2002; Agneessens *et al.*, 2004) and a particularly original approach based on boolean algebra in Bonacich, 1978.

Therefore, there already exists quite an impressive amount of work on 2-mode networks, and on methods for their analysis. However, we observe that many of the approaches proposed previously, though very relevant, are hardly applicable to *large* networks, typically networks with several hundreds of thousands nodes. Moreover, they often rely on quite complex notions and formalisms, which are difficult to handle for people only interested in analysing a given network. Finally, none of them consists in a generalisation of the post-1998 notions outlined in Section 1,

which are nowadays widely used to analyse 1-mode networks.

We propose here such a contribution. We design simple notions and methods to analyse very large 2-mode networks, which could be used as a first step in particular studies. These methods may then be extended to fit the details of particular cases, and we explain how to do so. Moreover, they are not only extensions of classical notions; we go further by proposing new notions designed specifically for the bipartite case. Our approach may also be applied to smaller networks, as long as they are not too small (typically thousands of nodes).

As explained above, the topic has a deep interdisciplinary nature. In order to make our techniques usable by a wide audience, we give a didactic presentation and we focus on basic notions. Let us insist however on the fact that this presentation is rigorous and formal, and, as will appear all along the paper, the results are sufficient to bring a significant amount of information on a given network.

Finally, we insist on the fact that analysing properly and in details a given network is a difficult task, which may be handled using different methods. There is no unique way to obtain relevant information and results in such cases. Moreover, much resides in the interpretations made from the outputs of these approaches. All the ones we have cited above, and the one we propose here, should therefore be seen as complementary rather than concurrent.

Let us conclude this section by noticing that, because of the wide dispersion of contributions due to the interdisciplinary nature of the topic (and the fact that it received continuous attention since several decades), we certainly missed some references. We however expect that the ones we have cited span well the contributions on the topic.

5 Methodology and data.

As already said, the methodology we follow has mainly been developped since the publication of the seminal paper Watts & Strogatz, 1998, and thus we call it the post-1998 approach. It relies on the introduction of statistical parameters aimed at capturing a given feature of networks under concern, and then on the comparison of the behaviours of real-world networks concerning these parameters as compared to random ones [8]. The underlying principle is that a parameter which behaves similarly on real-world and random networks is just a property of *most* networks (of which random networks are representatives) and so, though it may play an important role, it should not be considered as surprising and meaningful concerning the description of the real-world network. Instead, one generally looks for properties which make real-world networks different from most networks.

Our contribution here relies on this methodology. Namely, we will define statistical parameters aimed at capturing properties of bipartite graphs, and then evaluate the relevance of these parameters by comparing their values on random bipartite graphs and on real-world 2-mode networks.

[8] In the whole paper, the term *random* refers to object chosen uniformly at random in the given class: every element of the class has the same probability to be chosen. For descriptions on how to generate such graphs, we refer to Erdös & Rényi, 1959; Bollobas, 2001; Newman *et al.*, 2001a; Guillaume & Latapy, 2004b; Viger & Latapy, 2005.

Just like one considers purely random graphs and random graphs with prescribed degree distributions in the case of 1-mode networks, we will use both purely random bipartite graphs and random bipartite graphs with prescribed degree distributions. Such graphs are constructed easily by extending the 1-mode case, see for instance Newman *et al.*, 2001a; Guillaume & Latapy, 2004b [9]. Note that these models (both the 1-mode and 2-mode versions) generate graphs that are not necessarily *simple*: they may contain some loops and multiple links. There are however very few such links, and simply removing them generally has no impact on the results. This is what is generally done in the literature, and we will follow this convention here: in our context, it cannot have a significant impact [10].

Notice also that the properties of random graphs may be formally studied, see for instance Newman *et al.*, 2001a; Guillaume & Latapy, 2004a. One may also evaluate the mean properties of these graphs, and their standard deviations, using typically approaches like the ones developped in the p-star or ERGM (exponential random graph models) frameworks (*e.g.*, Robins *et al.*, 2005) [11]. However, our purpose here is only to identify properties that make real-world data different from random ones, not to quantify these differences precisely. We will therefore only compare empirical data to a typical random graph of the considered class (the fact that it is typical was checked by reproducing many times our experiments, which led to the same observations), and leave these investigations for further work, see Section 10.

In order to complete our comparison between random and real-world cases, we also need a set of real-world 2-mode networks. We chose the following four instances, which correspond to the examples given in the introduction and have the advantage of spanning well the variety of cases met in practice:

- the *actors-movies* network as obtained from the *Internet Movie Data Base* [12] in 2005, concerning $n_\perp = 127{,}823$ actors and $n_\top = 383{,}640$ movies, with $m = 1{,}470{,}418$ links;
- an *authoring* network obtained from the online *arXiv* preprint repository [13], with $n_\top = 19{,}885$ papers, $n_\perp = 16{,}400$ authors, and $m = 45{,}904$ links;
- an *occurrence* graph obtained from a version of the Bible [14] which contains $n_\perp = 9{,}264$ words and $n_\top = 13{,}587$ sentences with $m = 183{,}363$ links;
- a *peer-to-peer* exchange network obtained by registering all the exchanges processed by a large server during 48 hours (Guillaume *et al.*, 2005; Guillaume *et al.*, 2004), leading to $n_\top = 1{,}986{,}588$ peers, $n_\perp = 5{,}380{,}546$ data, and $m = 55{,}829{,}392$ links;

[9] We provide a program generating such graphs at `http://jlguillaume.free.fr/www/programs.php`

[10] One may also use the methods described in Viger & Latapy, 2005 to obtain directly simple (connected) graphs, but this is more intricate, and unnecessary in our context.

[11] See `http://www.sna.unimelb.edu.au/pnet/pnet.html` and `http://csde.washington.edu/statnet/`.

[12] See `http://www.imdb.com/`.

[13] See `http://arxiv.org/`.

[14] See `http://www.tniv.info/bible/`.

We provide these data, together with the programs computing the statistics described in this paper [15]. The key point here is that this dataset spans quite well the variety of context in which large 2-mode networks appear, as well as the variety of data sizes.

Let us insist on the fact that our aim here is not to derive conclusions on these particular networks: we only use them as real-world instances to illustrate the use of our results and to discuss their generality. This is why we do not detail more the way they are gathered and their relevance to any study. This is discussed in various references and is out of the scope of this paper.

6 Basic bipartite statistics.

The basic statistics on bipartite graphs are direct extensions of the ones on classical (1-mode) graphs. One just has to be careful with the fact that some classical properties give birth to twin bipartite properties while others must be redefined.

Let us consider a bipartite graph $G = (\top, \bot, E)$. We denote by $n_\top = |\top|$ and $n_\bot = |\bot|$ the numbers of top and bottom nodes, respectively. We denote by $m = |E|$ the number of links in the network. This leads to a top average degree $k_\top = \frac{m}{n_\top}$ and a bottom one $k_\bot = \frac{m}{n_\bot}$. One may obtain the average degree in the graph $G' = (\top \cup \bot, E)$ as $k = \frac{2m}{n_\top + n_\bot} = \frac{n_\top k_\top + n_\bot k_\bot}{n_\top + n_\bot}$. Finally, we obtain the bipartite density $\delta(G) = \frac{m}{n_\top n_\bot}$, *i.e.* the fraction of existing links with repect to possible ones. Note that this is different from the density of G': $\delta(G') = \frac{2m}{(n_\top + n_\bot)(n_\top + n_\bot - 1)}$, which is much lower.

Concerning the average distance (again, we restrict distance computations [16] to the largest connected component (denoted by *lcc*), which contains the vast majority of nodes, see Table 2), there is no crucial difference except that one may be interested by the average distance between top nodes and between bottom nodes, $d_\top$ and $d_\bot$. These values may be significantly different but one may expect that they are very close since a path between two top (resp. bottom) nodes is nothing but a path between bottom (resp. top) nodes with two additionnal links. Notice that there is no simple way to derive the average distance d in G' from the bipartite statistics $d_\bot$ and $d_\top$.

The values obtained for each of these basic properties on our four examples, together with values obained for random bipartite networks with the same size, are given in Table 2. It appears clearly that our examples may be considered as large networks with small average degrees, compared to their size. The density therefore is small. Moreover, the average distance is also small. These basic properties are very similar to what is observed on 1-mode networks: both 1-mode and 2-mode large real-world networks are sparse and have a small average distance, and in both contexts this is also true on random graphs.

[15] See `http://www.liafa.jussieu.fr/~latapy/Bip/`.

[16] Distance computations are expensive; the exact value cannot be computed in a reasonable amount of time for data of the size we consider here. Instead, we approximate the average by computing the average distance from a subset of the nodes to all the others, this subset being large enough to ensure that increasing it does not improve our estimation anymore, which is a classical method. All other computations are exact.

	actors-movies		authoring		occurrences		peer-to-peer	
	real	random	real	random	real	random	real	random
$n_\top$	127,823	*idem*	19,885	*idem*	13,587	*idem*	1,986,588	*idem*
$n_\bot$	383,640	*idem*	16,400	*idem*	9,264	*idem*	5,380,546	*idem*
m	1,470,418	*idem*	45,904	*idem*	183,363	*idem*	55,829,392	*idem*
$k_\top$	11.5	*idem*	2.3	*idem*	13.5	*idem*	28.1	*idem*
$k_\bot$	3.8	*idem*	2.8	*idem*	19.8	*idem*	10.4	*idem*
k	5.7	*idem*	2.5	*idem*	16.0	*idem*	15.2	*idem*
δ	0.000030	*idem*	0.00014	*idem*	0.0015	*idem*	0.0000052	*idem*
$lcc_\top$	124,414	125,944	16,209	18,512	13,579	13,587	1,986,343	1,426,978
$lcc_\bot$	374,511	381,431	11,654	14,607	9,246	9,264	5,380,507	5,054,689
$d_\top$	6.8	5.3	13.1	9.3	3.1	3.0	5.3	5.0
$d_\bot$	7.3	5.8	13.9	9.9	3.8	3.7	5.4	4.9
d	7.2	5.8	13.5	9.6	3.4	3.2	5.3	4.9

Table 2: Basic bipartite statistics on our four examples and on random bipartite graphs with the same size (same number of nodes and links, and thus same density and average degree as the real-world ones).

7 Bipartite statistics on degrees.

The notion of degree distribution has an immediate extension to the bipartite case. We denote by $\bot_i$ the fraction of nodes in $\bot$ having degree i and by $\top_i$ the fraction of nodes in $\top$ having degree i, and then call $(\bot_i)_{i\geq 0}$ the bottom degree distribution and $(\top_i)_{i\geq 0}$ the top one. See the appendix, page 30, for more detailed definitions and hints on how to understand this kind of statistics.

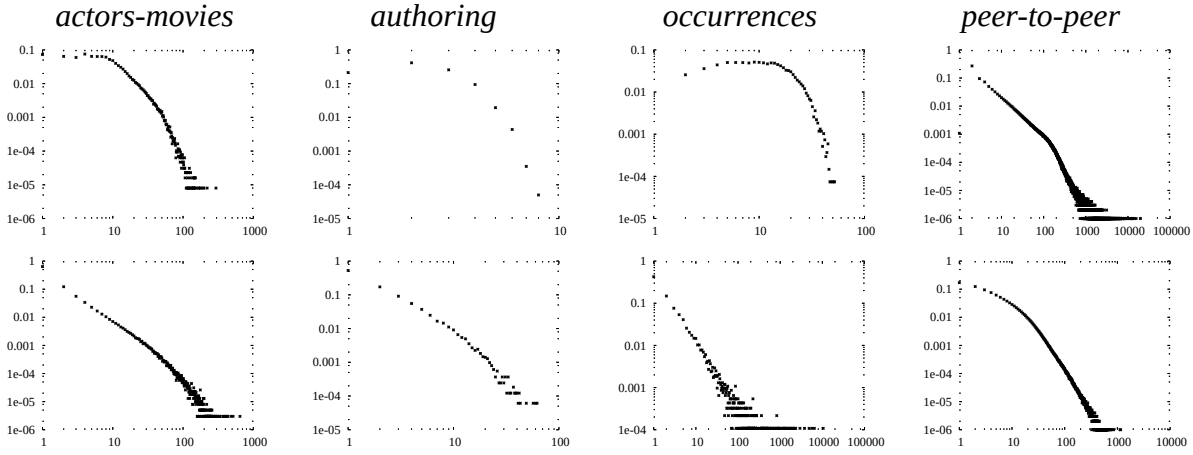

Figure 2: Degree distributions in our four real-world 2-mode networks. First row: for top nodes. Second row: for bottom nodes.

The top and bottom degree distributions of our four examples are given in Figure 2. One may observe on these plots that the bottom degree distributions are very heterogeneous and well fitted

by power laws (of various exponents). This is true in particular for the occurrences graph, which is a well known fact for a long time (Zipf, 1932): the frequency of occurrences of words in a text generally follows a particular kind of power law, named *Zipf* law. Instead, the shape of the top degree distribution depends on the case under concern: whereas it is well fitted by a power law in the peer-to-peer and actors-movies cases, it is far from a power law in the authoring and occurrences cases. This is due to the fact that papers have a limited number of authors (none has one hundred authors for instance), and likewise sentences have a limited number of words. Moreover, the number of very short sentences also is not huge. In these two cases, one can hardly conclude that the top degrees are very heterogeneous.

We finally conclude that, even if heterogeneity is present on at least one side of a 2-mode network, this is not generally true for both sides. This separates real-world 2-mode networks into two distinct classes, which should be taken into account in practice. This also confirms that considering the bipartite statistics brings significant information as compared to the projections, which exhibit power law degree distributions in all cases.

Let us now compare these real-world statistics with random graphs. If one generates purely random bipartite graphs of the same size as the ones considered here, the ($\top$ and $\bot$) degree distributions are Poisson laws. Therefore, the heterogenity of some degree distributions is not present, and even in the cases where the distributions are not very heterogeneous they do not fit the random case. We will therefore compare in the following our real-world 2-mode networks to random bipartite graphs with the same size and the same (top and bottom) degree distributions.

The next natural step is to observe possible correlations [17] between top and bottom degrees. In order to do this, we plot in Figure 3 the average degree of neighbours of nodes as a function of their degree, both for top and bottom nodes, separately. In other words, for each integer i we plot the average degree of all nodes which are neighbours of a node of degree i. We plot the same values obtained for random graphs of the same size and same degree distributions.

In the cases of actors-movies and peer-to-peer, the plots for the random cases are close to horizontal lines, showing that there are no correlations between a node degree and the average degree of its neighbours: this last value is independent of the node degree. In both cases, however, the real-world network displays nontrivial correlations. In the case of actors-movies, for instance, the average degree of neighbours of bottom nodes (the lower-left corner plot in Figure 3) decreases with the node degree. In other words, if an actor plays in many movies then he/she tends to play in smaller movies (in terms of the number of involved actors). Such nontrivial observations may be made on the other plots for actors-movies and peer-to-peer as well.

In the cases of authoring and occurrences, the plots for the random graphs are nontrivial: they grow for the top statistics, and are far from smooth for the bottom ones. Here again, the real-world cases exhibit significantly different behaviours, at least for the top statistics, thus demonstrating that these behaviours are nontrivial and related to intrinsinc properties of the underlying networks. Detailing this however is out of the scope of this paper. The key point here is to have evidence of the relevance of these statistics.

Notice that, despite they already bring much information, the statistics observed until now are almost immediate extensions of the classical ones. One may wonder if the bipartite nature of

[17] See the appendix, page 30 for more detailed definitions and hints on how to understand this kind of statistics.

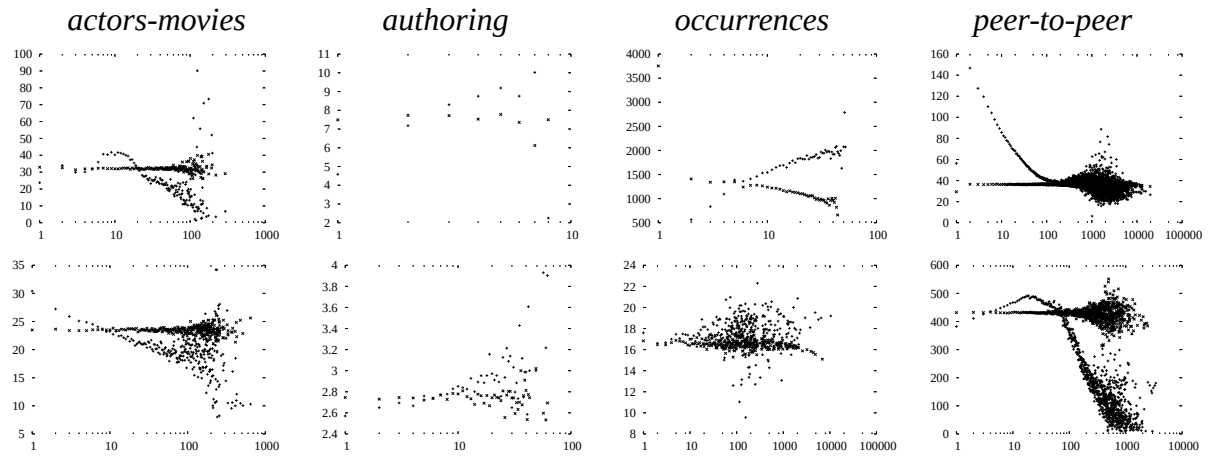

Figure 3: Degree correlations in our four real-world 2-mode networks, and in random bipartite graphs of the same size and same degree distributions. First row: for top nodes. Second row: for bottom nodes.

the networks under concern may lead to entirely new notions concerning degrees. We propose one below, with its variants.

Let us consider a node v in a bipartite graph $G = (\top, \bot, E)$, and let us denote by $N(N(v))$ the nodes at distance 2 from v, not including v, called *distance* 2 *neighbours* of v. We will suppose that v is a top node, the other case being dual. Notice that $N(N(v)) \subseteq \top$, and actually $N(N(v))$ is nothing but $N(v)$ in the $\top$-projection $G_\top$. The integer $|N(N(v))|$ therefore plays a central role in the projection approach, since it is the degree of v in $G_\top$.

But there are several ways for v to be linked to the nodes in $N(N(v))$, this information being lost during the projection. The two extreme cases occur when v is linked to only one node u in $\bot$, with $N(u) = N(N(v))$, or when v is linked to $|N(N(v))|$ nodes in $\bot$, each being linked to only one other node in $\top$. Of course, intermediate cases may occur, and the actual situation may be observed by plotting the correlations between the degree of nodes v, *i.e.* $|N(v)|$, and their number of distance 2 neighbours, $|N(N(v))|$. These statistics therefore offer a way to study how node degrees in the projection appear, and to distinguish between different behaviours. For instance, they make it possible to say if a given author has many coauthors because he/she writes many papers or if he/she writes papers with many authors. Such an information is not available in the projection of the authoring 2-mode network.

The plots in Figure 4 show that, as one may have guessed, the number of distance 2 neighbours of a node grows with its degree; more precisely, it generally grows as a power of the degree (the plots follow straight lines in log-log scale), and actually almost linearly. This is in conformance with the intuition that the number of distance 2 neighbours should be close to the degree of the node times the average degree of its neighbours. In the random cases, this leads to very straight plots (except in the top plot of occurrences). The real-world plots are quite close to the random ones, with a few notable exceptions: the slope of the plot is significantly different for the top plot of peer-to-peer, the real-world plots often are significantly below the random ones for large degrees, and they are in general slightly lower than the random ones even for small degrees.

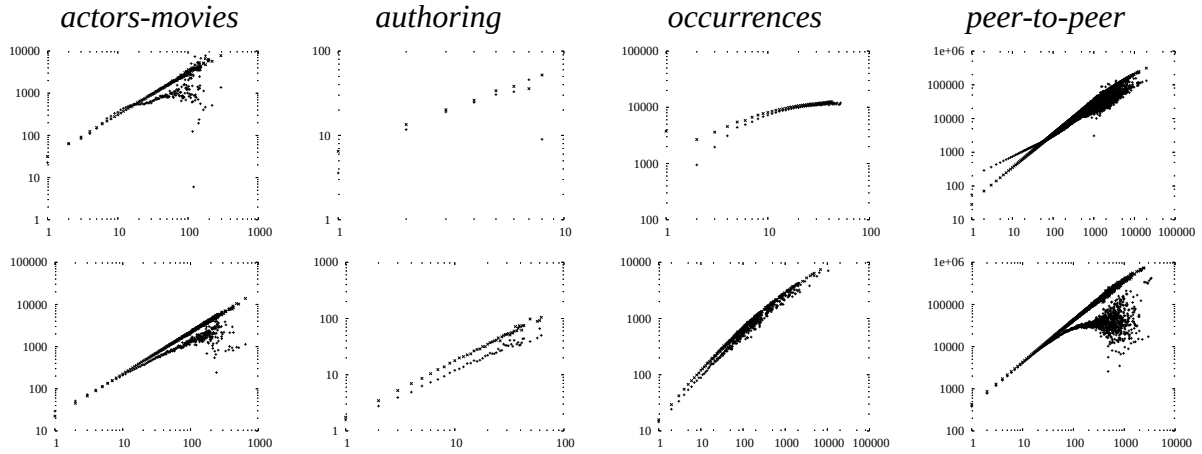

Figure 4: Correlations of the number of distance 2 neighbours with node degrees in our four examples, and in random bipartite graphs with the same size and degree distributions. First row: for top nodes. Second row: for bottom nodes.

This means that there is some redundancy in the neighbourhoods: whereas in random cases the number of distance 2 neighbours is close to the sum of the degrees of the direct neighbours, in real-world cases the direct neighbours have many neighbours in common and so the number of distance 2 neighbours is significantly lower. This is an important feature of large real-world networks, that we will deepen in the next sections.

8 Bipartite clustering and overlap.

Whereas there were quite direct extensions of the basic statistics and the ones on degrees to the bipartite case, the notion of clustering coefficient does not make any sense in itself in this context. Indeed, it relies on the enumeration of the triangles in the graphs, and there can be no triangle in a bipartite graph. We will therefore have to discuss the features captured by the classical clustering coefficients in order to propose bipartite extensions.

Both definitions of classical clustering coefficients capture the fact that when two nodes have something in common (one neighbour) then they are linked together with a probability much higher than two randomly chosen nodes. Conversely, they capture the fact that when two nodes are linked together then they probably have neighbours in common. In other words, they capture correlations between neighbourhoods. We will use this point of view here and define a first notion of clustering coefficient defined for pairs of nodes (in the same set $\top$ or $\bot$):

$$\mathrm{cc}_\bullet(u,v) = \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|}.$$

This is the most direct generalisation of the classical notion, and it was already suggested in Borgatti & Everett, 1997, and explicitely used in Guillaume *et al.*, 2005 in the context of peer-to-peer exchange analysis. It captures the overlap between neighbourhoods of nodes: if u and v

have no neighbour in common then $\text{cc}_\bullet(u,v) = 0$. If they have the same neighbourhood, then $\text{cc}_\bullet(u,v) = 1$. And if their neighbourhoods partially overlap then the value is in between, closer to 1 when the overlap is large compared to their degrees. See Figure 5 for an illustration.

This definition however has several drawbacks. The first one is the fact that it defines a value for *pairs* of nodes. One may want to capture the tendency of *one* particular node to have its neighbourhood included in the ones of other nodes. To achieve this, one may simply define the clustering coefficient of one node as the average of its clustering coefficients with other nodes. We however do not include in this averaging the pairs for which the overlap is empty [18]: most nodes have disjoint neighbourhood, which does not bring information. Like in the 1-mode case, we want to measure the implication of the fact of having one neighbour in common on the rest of the neighbourhoods. We finally obtain:

$$\text{cc}_\bullet(u) = \frac{\sum_{v \in N(N(u))} \text{cc}_\bullet(u,v)}{|N(N(u))|}$$

One may then observe the distribution of these values, their correlations with degrees, etc. One may also define the clustering coefficient of the top (resp. bottom) nodes, denoted by $\text{cc}_\bullet(\top)$ (resp. $\text{cc}_\bullet(\bot)$) as the average of this value over top (resp. bottom) nodes. The average over the all graph, denoted by $\text{cc}_\bullet(G)$, can then be obtained easily: $\text{cc}_\bullet(G) = \frac{n_\top \text{cc}_\bullet(\top) + n_\bot \text{cc}_\bullet(\bot)}{n_\top + n_\bot}$. We will discuss the obtained values below, see Table 3.

The notion of clustering coefficient discussed until now is an extension of the first classical one. It captures the fact that a node which has a neighbour in common with another node generally has a significant portion of neighbours in common with it. There is another way to capture this, similar to the second definition of classical clustering coefficient, is to measure the probability that, given four nodes with three links, they actually are connected with four links (all the possible bipartite ones):

$$\text{cc}_\mathsf{N}(G) = \frac{2N_\bowtie}{N_\mathsf{N}}$$

where $N_\bowtie$ is the number of quadruplets of nodes with four links in G, and N_N is the number of quadruplets of nodes with at least three. This extension of the second notion of classical clustering coefficient was already proposed in Robins & Alexander, 2004 in the context of company board networks. It is a natural generalisation of the clustering coefficient $cc_\vee$ on classical (1-mode) graphs: this last notion is the probability, when three nodes are linked in a chain (with two links), that they form a triangle; the cc_N notion is nothing but the probability, when four nodes are linked in a chain (with three links), that they form a square. This extension is natural since there cannot be any triangle in bipartite graphs. We will discuss the obtained values below, see Table 3.

The two notions above generalise the classical definitions of clustering coefficients. Capturing the overlap between neighbours may however need more precision. Suppose that degrees are heterogeneous in the network, as it is often the case (Section 7), and consider two nodes u and v.

[18] As a consequence, the obtained value will never be 0, but it may be very small. Notice also that the clustering coefficient is not defined for nodes v such that $N(N(v)) = \emptyset$ (recall that, by definition, $v \notin N(N(v))$.

If one of these nodes has a high degree and the other has not, then $\mathrm{cc}_{\bullet}(u,v)$ will necessarily be small. This will be true even if one of the neighbourhoods is entirely included in the other. One may however want to capture this, which can be done using the following definition:

$$\mathrm{cc}_{\underline{\bullet}}(u,v) = \frac{|N(u) \cap N(v)|}{\min\left(|N(u)|, |N(v)|\right)}.$$

One may define dually:

$$\mathrm{cc}_{\overline{\bullet}}(u,v) = \frac{|N(u) \cap N(v)|}{\max\left(|N(u)|, |N(v)|\right)}.$$

See Figure 5 for an illustration. These two notions, called min- and max-clustering, were introduced first in Guillaume *et al.*, 2005. The first one emphasises on the fact that small neighbourhoods may intersect significantly large ones; it is equal to 1 whenever one of the neighbourhoods is included in the other. The second one emphasises on the fact that neighbourhoods (both small or large ones) may overlap very significantly: it is 1 only when the two neighbourhoods are the same and it tends to decreases rapidly if the degree of one of the involved nodes increases. It captures the fact that nodes with *similar* degrees have high neighbourhood overlaps.

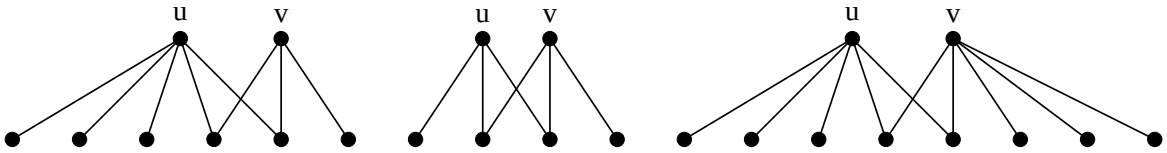

Figure 5: Examples of bipartite clustering coefficients, and interpretations. Left: a case in which $\mathrm{cc}_{\bullet}(u,v) = \frac{2}{6} = 0.333\cdots$ is quite small, despite the fact that u and v have two neighbours in common, due to the fact that the union of their neighbours is quite large; on the contrary, $\mathrm{cc}_{\underline{\bullet}}(u,v) = \frac{2}{3} = 0.666\cdots$ is quite large, revealing that one of the neighbourhoods is almost included in the other; the value of $\mathrm{cc}_{\overline{\bullet}}(u,v) = \frac{2}{5} = 0.4$ indicates that this may be due to the fact that one of the nodes has a high degree. The situation is different in the case at the center: all clustering coefficients are quite high (resp. 0.5, $0.666\cdots$, and $0.666\cdots$), indicating that there is not only an important overlap, but that this overlap concerns a significant part of each neighbourhoods (and thus the two nodes have similar degrees). On the right, the two nodes have a small clustering coefficient $\mathrm{cc}_{\bullet}(u,v) = \frac{2}{8} = 0.25$, and the fact that the value of $\mathrm{cc}_{\underline{\bullet}}(u,v) = \frac{2}{5} = 0.4$ remains quite small indicates that this is not due to the fact that one of the two nodes has a very high degree compared to the other one. If ones considers larger degree nodes, then the difference between *small* and *high* values is clearer, but the figure would be unreadable.

With these definitions, one may define $\mathrm{cc}_{\underline{\bullet}}(v)$, $\mathrm{cc}_{\underline{\bullet}}(\top)$, $\mathrm{cc}_{\underline{\bullet}}(\bot)$, $\mathrm{cc}_{\underline{\bullet}}(G)$, $\mathrm{cc}_{\overline{\bullet}}(v)$, $\mathrm{cc}_{\overline{\bullet}}(\top)$, $\mathrm{cc}_{\overline{\bullet}}(\bot)$, and $\mathrm{cc}_{\overline{\bullet}}(G)$ in a way similar to the one used above for $\mathrm{cc}_{\bullet}(v)$, $\mathrm{cc}_{\bullet}(\top)$, $\mathrm{cc}_{\bullet}(\bot)$, and $\mathrm{cc}_{\bullet}(G)$. The distributions and various correlations may then be observed.

We give in Table 3 the values obtained for our four examples together with the values obtained for random bipartite graphs with same size and degree distributions (the values for purely random bipartite graphs are similar). It appears clearly that the notions we introduced capture different kinds of overlaps between neighbourhoods. However, except for $\mathrm{cc}_N(G)$, the obtained values

	actors-movies		authoring		occurrences		peer-to-peer	
	real	random	real	random	real	random	real	random
$cc_\bullet(\top)$	0.064	0.046	0.29	0.27	0.066	0.066	0.056	0.019
$cc_\bullet(\bot)$	0.36	0.20	0.31	0.25	0.065	0.038	0.076	0.074
$cc_N(G)$	0.0082	0.00024	0.079	0.00012	0.053	0.048	0.0094	0.00019
$cc_{\underline{\bullet}}(\top)$	0.24	0.23	0.56	0.56	0.19	0.20	0.27	0.24
$cc_{\underline{\bullet}}(\bot)$	0.81	0.79	0.73	0.70	0.64	0.61	0.39	0.42
$cc_{\overline{\bullet}}(\top)$	0.087	0.062	0.36	0.34	0.097	0.097	0.074	0.024
$cc_{\overline{\bullet}}(\bot)$	0.37	0.21	0.33	0.26	0.069	0.041	0.091	0.089

Table 3: Bipartite clustering statistics on our four examples and on random bipartite graphs with the same size and same degree distributions.

are not very different on random graphs and on real-world networks. This indicates that these statistics do not capture a very significant feature of large real-world networks, which will discuss this further below. Instead, the obtained values for $cc_N(G)$ is significantly larger on real-world networks than on random graphs, which shows that it captures more relevant information.

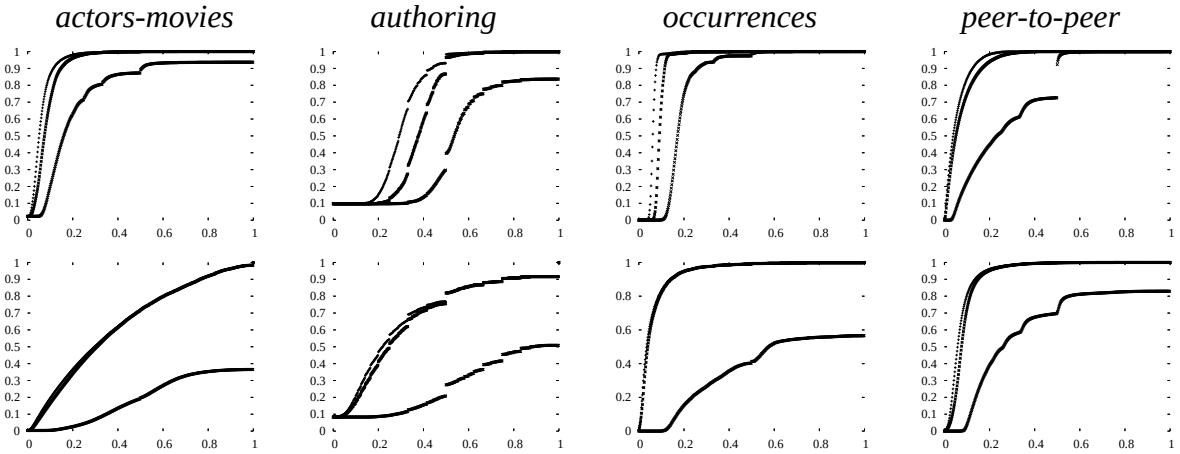

Figure 6: Cumulative distributions of the various clustering coefficients in our four real-world 2-mode networks. First row: for top nodes. Second row: for bottom nodes.

We show in Figure 6 the cumulative distributions [19] of $cc_\bullet(v)$, $cc_{\underline{\bullet}}(v)$ and $cc_{\overline{\bullet}}(v)$ for our four examples, *i.e.* for each value x on the horizontal axis the ratio of all the nodes having a value lower than x for these statistics. Before entering in the discussion of these plots, notice that, by definition, we have $cc_\bullet(v) \leq cc_{\overline{\bullet}}(v) \leq cc_{\underline{\bullet}}(v)$ for any v. Therefore, the lower plots in each case of Figure 6 is the one of $cc_{\underline{\bullet}}(v)$, the upper is the one for $cc_\bullet(v)$ and the one for $cc_{\overline{\bullet}}(v)$ is in between.

More interesting, the plots exhibit quite different behaviours. In several cases (in particular top of actors-movies, occurrences and peer-to-peer, as well as bottom of occurrences and peer-

[19] See the appendix, page 30 for more detailed definitions and hints on how to understand this kind of statistics.

to-peer) the plots for $\text{cc}_{\overline{\bullet}}(v)$ and $\text{cc}_{\bullet}(v)$ grow very rapidly and are close to 1 almost immediately. This means that the values of these statistics are very small, almost 0, for most nodes: in these cases, the neighbours of nodes have a small intersection, compared to the union of their neighbourhoods. However, in several cases, the plots for $\text{cc}_{\underline{\bullet}}(v)$ grow much less quickly, and remain lower than 1 for a long time. In several cases, it is even significantly lower than 1 by the end of the plot, meaning that for an important number of nodes the value of $\text{cc}_{\underline{\bullet}}(v)$ is equal to 1: almost 10% in the case of top of actors-movies, almost 20% in the cases of top authoring and bottom of peer-to-peer, and more than 40% in the case of bottom of occurrences. This means that, despite overlaps are in general small compared to their possible value, the neighbourhoods of many low-degree nodes significantly or even completely overlap with other nodes neighbours.

Other cases display a very different behaviour: in both top and bottom plots of authoring, and in bottom of actors-movies, it appears clearly that a significant number of nodes have a large value for $\text{cc}_{\underline{\bullet}}(v)$, $\text{cc}_{\bullet}(v)$ and $\text{cc}_{\overline{\bullet}}(v)$. This means that node neighbours overlap significantly, and that this is not only a consequence of the fact that low degree nodes have their neighbourhoods included in the ones of other nodes.

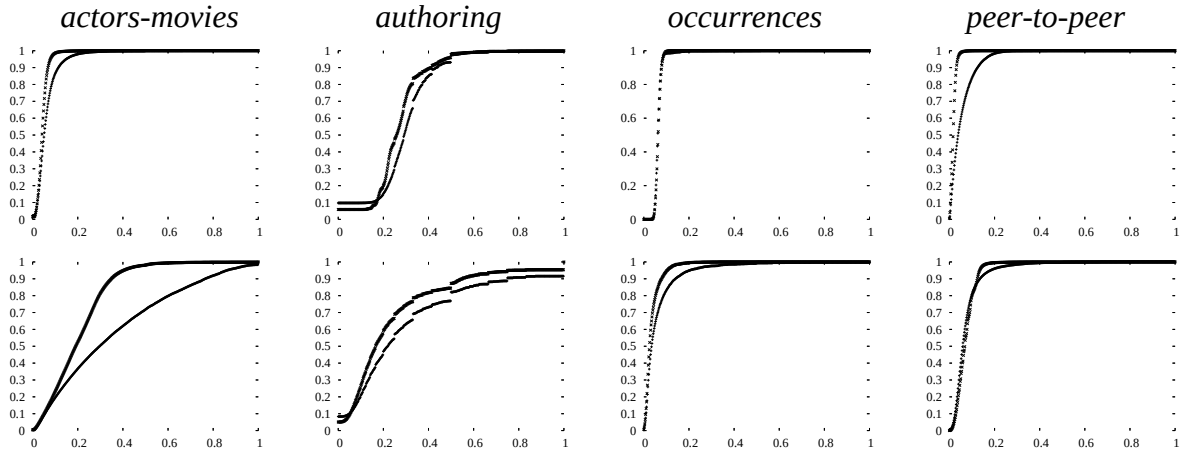

Figure 7: Cumulative distributions of the $\text{cc}_{\bullet}$ clustering coefficient in our four real-world 2-mode networks, and in random bipartite graphs of the same size and same degree distributions. First row: for top nodes. Second row: for bottom nodes.

Again, our aim here is not to discuss in detail the specificities of each case, but to give evidence of the fact that these statistics have nontrivial behaviours and capture significant information. It is clear from the discussion above that the three notions of clustering captured by $\text{cc}_{\underline{\bullet}}(v)$, $\text{cc}_{\bullet}(v)$ and $\text{cc}_{\overline{\bullet}}(v)$ are different, and give complementary insight on the underlying network properties. One may however be surprised by the fact that $\text{cc}_{\bullet}(v)$ often is very small, which we deepen now by comparing its behaviours on real-world cases and on random ones, see Figure 7 [20].

[20] For clarity and to avoid long discussions on specific behaviours, which is out of our scope here, we only compare the real-world and the random behaviours of $\text{cc}_{\bullet}(v)$ (not of the two other notions of clustering coefficients).

In these plots, it appears clearly that, except in the case of bottom of actors-movies, the plots of the real-world values and of the random ones are quite similar. This means that, concerning the values of $\mathrm{cc}_\bullet(v)$, real-world graphs are not drastically different from random ones (they however have slightly higher values of $\mathrm{cc}_\bullet(v)$ in most cases). In other words, this statistics does not capture very significant information, according to the methodology described in Section 5. This is due to the fact that the low degree nodes (which are numerous in our networks) have with high probability their neighbours in common with high degree nodes; by definition, this induces a low value for $\mathrm{cc}_\bullet(v)$, and even lower for $\mathrm{cc}_{\overline{\bullet}}(v)$. This is true by construction for random graphs, and the plots above show that this is mostly true for real-world networks also, which was not obvious.

Similar conclusions follow from the study of $\mathrm{cc}_{\overline{\bullet}}(v)$, but the study of $\mathrm{cc}_{\underline{\bullet}}(v)$ leads to the opposite conclusion: an important number of nodes have their neighbourhood included in the one of other (large degree) nodes, as already discussed, which happens much more rarely in random graphs. We do not detail these results here, since they do not fit in the scope of this paper. Instead, we will propose a new statistics in the next section that has several advantages on the clustering coefficients discussed here and does not have their drawbacks.

Before turning to this other statistics, let us observe the correlations between node degrees and their clustering coefficient. Again, for clarity and to maintain the paper within a reasonable length, we focus on $\mathrm{cc}_\bullet(v)$ and its comparison with the random case. See Figure 8.

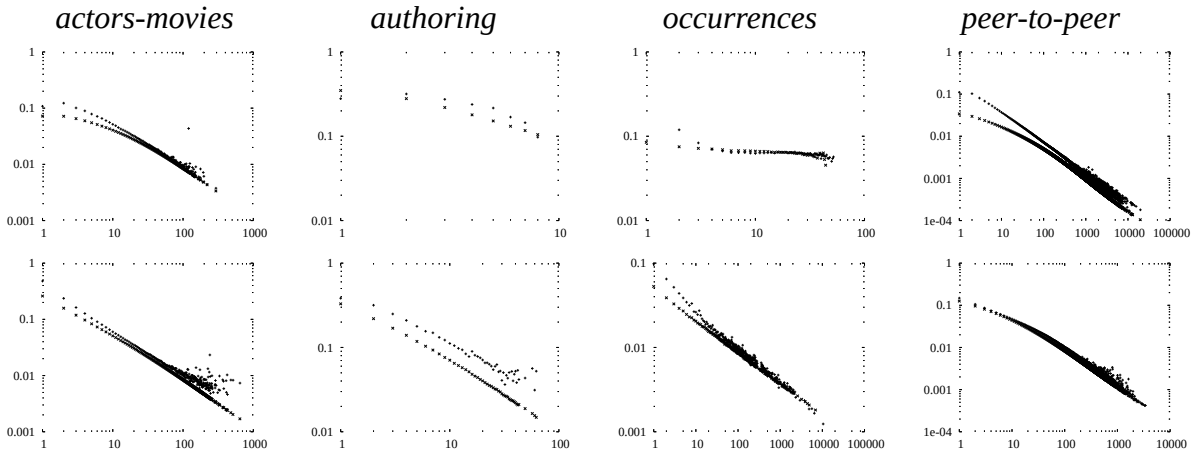

Figure 8: Correlations of the $\mathrm{cc}_\bullet(v)$ clustering coefficient with node degrees in our four examples, and in random bipartite graphs with the same size and degree distributions. First row: for top nodes. Second row: for bottom nodes.

The values for the random graphs are below the ones for the real-world cases (or they coincide at some points), in all plots. This shows that the value of $\mathrm{cc}_\bullet(v)$ are larger in real-world cases than in random ones, but the difference is small, which confirms the observations above. More interestingly, it appears clearly that in most cases $\mathrm{cc}_\bullet(v)$ decreases as a power of the degree of v (straight line in log-log scale). In other words, the clustering coefficient of low degree nodes is quite large, but the one of large degree nodes is very small, like in random graphs.

9 The notion of redundancy.

In the previous section, we discussed several ways to extend the classical notions of clustering coefficient to the bipartite case. One may wonder if the bipartite nature of the networks under concern may lead to new, specific notions, just like we observed concerning degrees in Section 7. Moreover, one may want to capture the notion of overlap concerning *one* particular node; in previous section, this was only possible by averaging the value obtained for a possibly large number of pairs of nodes. This section answers this: it is devoted to a new notion aimed at capturing overlap in bipartite networks, in a node-centered fashion.

First notice that neighbourhood overlaps correspond to links which are obtained in several ways during the projection, and that these links cannot be distinguished one from another in the projection. They also reveal the fact that, among all the links induced by a node of a bipartite graph in the projection, many (and possibly all) may actually be induced by others too. In other words, if we remove this node from the bipartite graph then the projection may be only slightly changed (or even not at all). This can be captured by the following parameter, which we call the *redundancy coefficient* of v:

$$\mathrm{rc}(v) = \frac{|\{\{u,w\} \subseteq N(v),\ \exists\, v' \neq v,\ (v',u) \in E \text{ and } (v',w) \in E\}|}{\frac{|N(v)|(|N(v)|-1)}{2}}.$$

In other words, the redundancy coefficient of v is the fraction of pairs of neighbours of v linked to another node than v. In the projection, these nodes would be linked together even if v were not there, see Figure 9; this is why we call this the redundancy. If it is equal to 1 then the projection would be exactly the same without v; if it is 0 it means that none of its neighbours would be linked together in the projection [21].

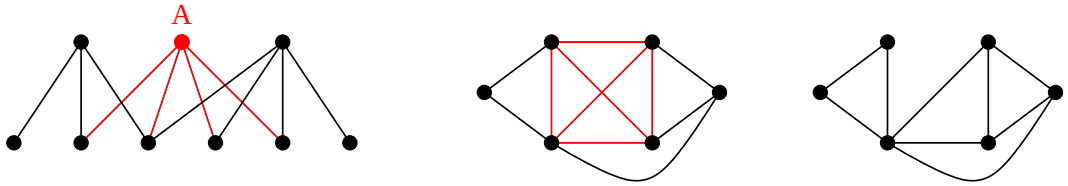

Figure 9: Example of redundancy computation. From left to right: a bipartite graph, its $\perp$-projection, and the $\perp$-projection obtained if the node A is first removed. Only two links disappear, leading to $\mathrm{rc}(A) = \frac{4}{6} = 0.666\cdots$.

Again, we can derive from this definition the ones of $\mathrm{rc}(\top)$, $\mathrm{rc}(\perp)$ and $\mathrm{rc}(G)$, as well as distributions and correlations. We give in Table 4 the values obtained for our four examples and for comparable random graphs. It appears clearly from these values that, except in the case of occurrences, the redundancy coefficient is much larger in real-world networks than in random

[21] Interestingly, the notion of redundancy we propose here is equivalent to the generalisation of the notion of clustering coefficient to squares, denoted by $C_4()$, proposed independently in Lind *et al.*, 2005: it is the probability, when a node has two neighbours, that these two nodes have (another) neighbour in common. Though the two points of view are quite different, and the definitions termed differently, the two notions are exactly the same.

graphs, and that it actually is very large: in peer-to-peer, for instance, on average half the pairs of peers that have a common interest for a given data also have a common interest for another data. These values are much larger than the ones for the clustering coefficients in the previous section, see Table 3, and the difference they make between random graphs and real-world networks is much more significant. To this regard, it may be considered as a better generalisation of clustering coefficients in 1-mode networks than the bipartite clustering coefficients defined in Section 8.

The case of occurrences is different: the projections on both sides are very dense, which is very particular as already noticed. The redundancy coefficient therefore is huge, but this is not because of a property of how the neighbourhoods overlap: this is a direct consequence of the high density of the projections. In such a case, the redundancy coefficient is meaningless, and we will therefore not discuss this case any further in this section; simply notice that the redundancy coefficient has similar behaviours in such graphs and in their random equivalent.

	actors-movies		authoring		occurrences		peer-to-peer	
	real	random	real	random	real	random	real	random
rc($\top$)	0.26	0.014	0.38	0.0016	0.80	0.74	0.31	0.011
rc($\bot$)	0.25	0.011	0.33	0.00037	0.83	0.75	0.50	0.069

Table 4: The redundancy coefficient for our four examples and for random bipartite graphs with the same size and same degree distributions.

We show in Figure 10 the distributions of $\mathrm{rc}(v)$ for our four examples together with plots for comparable random graphs. These plots confirm that the redundancy coefficient captures a property that makes large real-world networks different from random ones: in all the cases except occurrences, the value of this coefficient in random graphs is almost 0 for all nodes (both top and bottom); instead, in real-world networks it is significantly larger, and equal to 1 for a large portion of the nodes. This last fact is not surprising since $\mathrm{cc}_{\underline{\bullet}}(v) = 1$ implies $\mathrm{rc}(v) = 1$ for all nodes v.

However, the redundancy coefficient has a much wider range of values than $\mathrm{cc}_{\underline{\bullet}}(v)$, which generally is close to 0 or 1, see Figure 6. Moreover, the redundancy coefficient captures a different property: in the case of actors-movies, for instance, it does not only mean that a significant number of movies have a cast that is a sub-cast of another movie (as captured by $\mathrm{cc}_{\underline{\bullet}}(v)$), but that when two actors act together in a movie then there often exists (at least) another movie in which they also act together. Both are interesting, and complementary, but the redundancy coefficient certainly captures a more general feature.

Let us now observe the correlations between node redundancy coefficient and their degree, plotted in Figure 11. In these plots, except for occurrences, the plots for the random graphs coincide with the x-axis, which confirms that the values of node redundancy in random graphs are very small, independently of node degrees. Real-world cases, on the contrary, exhibit nontrivial behaviours. In most cases, the redundancy decreases with the degree, which is not surprising since the number of links needed in the projection in order for the redundancy of a node to be large grows with the square of its degree. However, the redundancy remains large even for quite

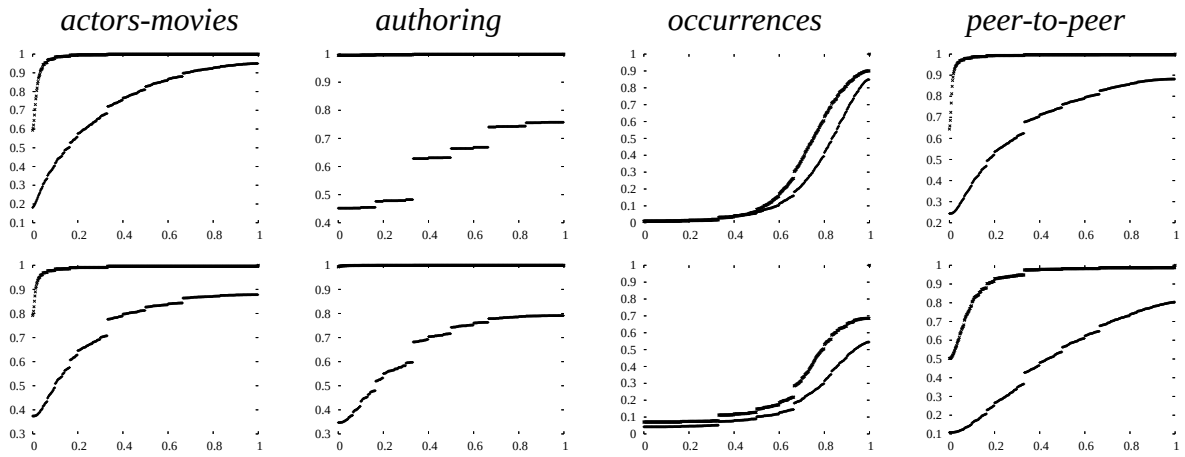

Figure 10: Cumulative distributions of the redundancy coefficient in our four real-world 2-mode networks, and in random bipartite graphs of the same size and same degree distributions. First row: for top nodes. Second row: for bottom nodes.

large degrees: it is close to 0.15 for nodes of degree 30 for top nodes in actors-movies, for instance, meaning that among the 435 possible pairs of neighbours of these nodes, on average 65 are linked to another top node in common. This has a very low probability in random graphs. Likewise, one may notice that some very high degree nodes have a very large redundancy coefficient in several cases, which also is a significant information.

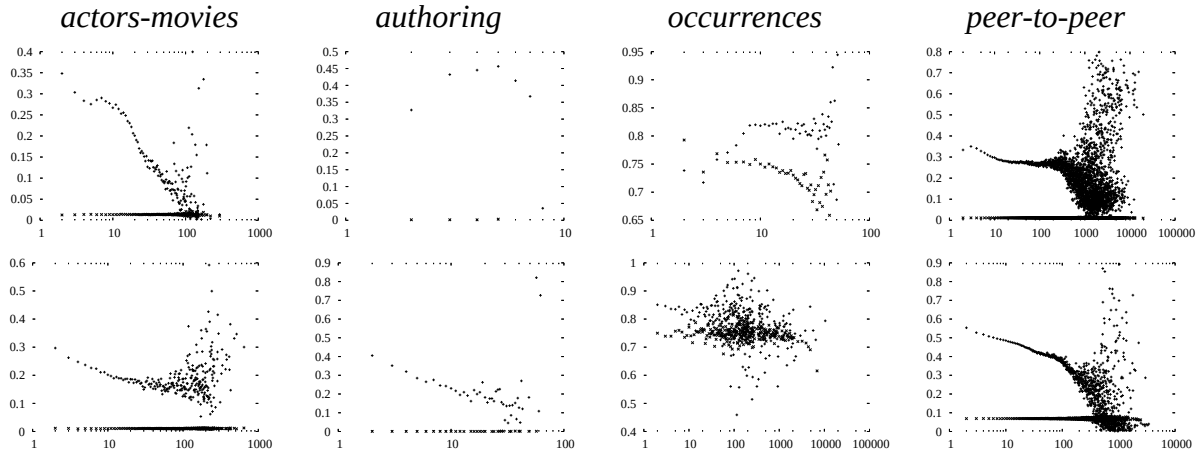

Figure 11: Correlations of redundancy coefficient with node degrees in our four real-world 2-mode networks, and in random bipartite graphs of the same size and same degree distributions. First row: for top nodes. Second row: for bottom nodes.

One may push further the study of the redundancy, for instance by counting how many nodes have an overlap with a given one, and so may be responsible for its high redundancy. This is nothing but the degree of the node in the appropriate projection, which emphasises once again

that our approach may be fruitfully combined with the one based on projection, as argued in Section 3.

10 Conclusion and perspectives.

The core contribution of this paper is a set of rigorous and coherent statistical properties usable as a basis for the analysis of large real-world 2-mode networks following the post-1998 approach. These statistics go from the very basics (size, distances, etc) to subtle ones (typically various clustering coefficients and their correlations with degrees). Let us insist on the fact that we do not only extend classical notions to the bipartite case, but also develop new notions which make sense only in this context. Moreover, the proposed approach avoids projection of 2-mode networks into 1-mode ones, which makes it possible to grab much richer information. We hope that this unified framework and discussion will help significantly people involved in analysis of such networks.

A first conclusion drawn from the computation of these statistics on four representative real-world examples is that, just like large real-world 1-mode networks, they have nontrivial properties in common which make them very different from random networks. In particular, there is a high heterogeneity between degrees of nodes of at least one kind, and there are significant overlaps between neighbourhoods. Concerning this last property, we show that immediate extensions of the classical notions of clustering coefficients are not sufficient to make the difference between real-world networks and random graphs; we propose the notion of *redundancy* as a relevant alternative. Overall, these facts are strikingly close to what is met in 1-mode networks and should play a similar role. Conversely, we observed many properties which behave differently depending on the 2-mode network under concern, which may be used to describe a particular instance in more details.

Notice that these contributions do not only concern the 2-mode networks themselves, but also their projection: keeping the bipartite nature of the data makes it possible to obtain more precise information on the projection itself. For instance, statistics on degrees make it possible to separate high degree nodes in the projection into two distinct classes: the ones which are linked to many nodes in the 2-mode network, and the ones linked to nodes of high degree in the 2-mode network. This kind of analysis could be deepened using clustering and redundancy notions.

Going further, one may use the notions we introduced here to define new relevant statistics on 1-mode networks. Indeed, any graph $G = (V, E)$ may be seen as a bipartite graph $G' = (V, V, E)$ where the links are between two *copies* of V. The statistics we studied here may then be computed on this bipartite graph, leading to new insight on the original graph G.

There are many directions to improve and continue the work presented here. Among them, let us cite the analytic study of the parameters we propose, which can in particular be done using the techniques in Newman *et al.*, 2001b or in Robins *et al.*, 2005. One might prove in this way the expected behaviour of these parameters and deepen their understanding. Another direction is the developement of models of 2-mode networks capturing the properties met in practice. Just as is the case for 1-mode networks, much can be done concerning degrees, see Newman *et al.*, 2001a; Guillaume & Latapy, 2004a, but very little is known concerning the

modeling of clustering and redundancy. Finally, applying these results to practical cases and giving precise interpretations of their meanings in these different contexts would probably help in designing other relevant notions. To this regard, the statistical properties described in this paper may help in deepening the key questions about group formation and relations (like the emergence of interlocking in company boards, see Robins & Alexander, 2004; Conyon & Muldoon, 2004; Battiston & Catanzaro, 2004; Newman *et al.*, 2001a or of scientific areas and communities, see Roth & Bourgine, 2005; Morris & Yen, 2005; Newman, 2001a; Newman, 2001b; Newman, 2000), which we did not consider here.

Let us conclude by noticing that the field of large network analysis is only at its beginning, though much has been done, before and after 1998, on 1-mode networks. However, most real-world networks are directed, weighted, labelled, hybrid, and/or evolve during time. Some work has recently been done concerning weighted networks (Barrat *et al.*, 2004; Barthélemy *et al.*, 2005; Newman, 2004), and we propose here a contribution concerning 2-mode networks. However, there is still much to do to be able to analyse efficiently these various kinds of networks. Extending the notions we discussed here to the case of multipartite graphs (nodes are in several disjoint sets, with links between nodes in different sets only) would be a step further in this direction.

Acknowledgments

We warmly thank Arnaud Bringé, Dominique Cardon, Pascal Cristofoli, Nicolas Gast, Jean-Loup Guillaume, Christophe Prieur, Camille Roth and Fabienne Venant, as well as the *SocNet* community, for their precious comments and help during the preparation of this contribution. We also thank the anonymous reviewers and Professor Patrick Doreian, who managed the submission to *Social Networks*, for the great work they did and their help in improving the initial version. This work was funded in part by the PERSI (*Programme d'Étude des Réseaux Sociaux de l'Internet*) project and by the AGRI (*Analyse des Grands Réseaux d'Interactions*) project.

References

Agneessens, Filip, Roose, Henk, & Waege, Hans. 2004. Choices of Theatre Events: p* Models for Affiliation Networks with Attributes. *Metodoloski zvezki*, **1**(2), 419–439. Short version presented at SunBelt 2002.

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

Barabasi, A.-L., & Albert, R. 1999. Emergence of scaling in random networks. *Science*, **286**, 509–512.

Barrat, A., Barthélemy, M., Pastor-Satorras, R., & Vespignani, A. 2004. The architecture of complex weighted networks. *Proc. Natl. Acad. Sci. USA*, **101**.

Barthélemy, M., Barrat, A., Pastor-Satorras, R., & Vespignani, A. 2005. Characterization and Modeling of weighted networks. *Physica A*, **346**.

Battiston, Stefano, & Catanzaro, Michele. 2004. Statistical properties of Corporate Board and Director Networks. *European Physics Journal B*, **38**, 345–352.

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

Bollobas, B. 2001. *Random Graphs*. Cambridge University Press.

Bonacich, Phillip. 1972. Technique for Analyzing Overlapping Memberships. *Sociological Methodology*, **4**, 176–185.

Bonacich, Phillip. 1978. Using Boolean Algebra to Analyze Overlapping Memberships. *Sociological Methodology*, **9**, 101–115.

Borgatti, Stephen P., & Everett, Martin G. 1992. Regular blockmodels of multiway, multimode matrices. *Social Networks*, **14**, 91–120.

Borgatti, Stephen P., & Everett, Martin G. 1997. Network analysis of 2-mode data. *Social Networks*, **19**(3), 243–269.

Bornholdt, S., & Schuster, H.G. (eds). 2003. *Hankbook of Graphs and Networks: From the Genome to the Internet*. Wiley-VCH.

Boudourides, Moses A., & Botetzagias, Iosif A. 2004. Networks of Protest on Global Issues in Greece 2002-3. Work in progress – Preprint.

Brandes, Ulrik, & Erlebach, Thomas (eds). 2005. *Network Analysis: Methodological Foundations*. Lecture Notes in Computer Science, vol. 3418. Springer.

Breiger, Ronald L. 1974. The Duality of Persons and Groups. *Social Forces*, **53**(2), 181–190.

Caldarelli, Guido, Battiston, Stefano, Garlaschelli, Diego, & Catanzaro, Michele. 2004. Emergence of Complexity in Financial Networks. *Lecture Notes in Physics*, **650**, 399–423.

Conyon, Martin J., & Muldoon, Mark R. 2004. The Small World Network Structure of Boards of Directors. SSRN preprint, `http://ssrn.com/abstract=546963`.

Dahui, Wang, Li, Zhou, & Zengru, Di. 2005. Bipartite Producer-Consumer Networks and the Size Distribution of Firms. ArXiV preprint `physics/0507163`.

Dhillon, Inderjit S. 2001. Co-clustering documents and words using bipartite spectral graph partitioning. *Pages 269–274 of: KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM Press.

Doreian, Patrick, Batagelj, Vladimir, & Ferligoj, Anuska. 2004. Generalized blockmodeling of two-mode network data. *Social Networks*, **26**, 29–53.

Erdös, P., & Rényi, A. 1959. On random graphs I. *Publications Mathematics Debrecen*, **6**, 290–297.

Ergun, Guler. 2002. Human Sexual Contact Network as a Bipartite Graph. ArXiV preprint `cond-mat/0111323`.

Faloutsos, M., Faloutsos, P., & Faloutsos, C. 1999. On Power-law Relationships of the Internet Topology. *Pages 251–262 of: SIGCOMM*.

Faust, K. 2005. *Models and Methods in Social Network Analysis*. New York: Cambridge University Press. Chap. Using Correspondence Analysis for Joint Displays of Affiliation Networks.

Faust, Katherine. 1997. Centrality in affiliation networks. *Social Networks*, **19**, 157–191.

Faust, Katherine, Willert, Karin E., Rowlee, David D., & Skvoretz, John. 2002. Scaling and statistical models for affiliation networks: patterns of participation among Soviet politicians during the Brezhnev era. *Social Networks*, **24**, 231–259.

Ferrer, R., & Solé, R.V. 2001. The Small-World of Human Language. *Pages 2261–2265 of: Proceedings of the Royal Society of London*, vol. B268.

Fessant, F., Handurukande, S., Kermarrec, A.-M., & Massoulié, L. 2004. Clustering in Peer-to-Peer File Sharing Workloads. *In: 3-rd International workshop on Peer-To-Peer Systems (IPTPS)*.

Freeman, Linton C. 2003. Finding social groups: A meta-analysis of the southern women data.

Garlaschelli, Diego, Battiston, Stefano, Castri, Maurizio, Servedio, Vito D. P., & Caldarelli, Guido. 2004. The scale-free Topology of market investments. ArXiV preprint `cond-mat/0310503`.

Guillaume, Jean-Loup, & Latapy, Matthieu. 2004a. Bipartite graphs as Models of Complex Networks. *In: Lecture Notes in Computer Sciences (LNCS), proceedings of the 1-st International Workshop on Combinatorial and Algorithmic Aspects of Networking (CAAN)*.

Guillaume, Jean-Loup, & Latapy, Matthieu. 2004b. Bipartite Structure of *all* Complex Networks. *Information Processing Letters (IPL)*, **90**(5), 215–221.

Guillaume, Jean-Loup, Le Blond, Stevens, & Latapy, Matthieu. 2004. Statistical analysis of a P2P query graph based on degrees and their time-evolution. *In: Lecture Notes in Computer Sciences (LNCS), proceedings of the 6-th International Workshop on Distributed Computing (IWDC)*.

Guillaume, Jean-Loup, Le Blond, Stevens, & Latapy, Matthieu. 2005. Clustering in P2P exchanges and consequences on performances. *In: Lecture Notes in Computer Sciences (LNCS), proceedings of the 4-th international workshop on Peer-to-Peer Systems (IPTPS)*.

Holme, Petter, Park, Sung Min, Kim, Beom Jun, & Edling, Christofer R. 2004. Korean university life in a network perspective: Dynamics of a large affiliation network. ArXiV preprint `cond-mat/0411634`.

Iamnitchi, Adriana, Ripeanu, Matei, & Foster, Ian. 2004. Small-World File-Sharing Communities. *Proceedings of the 23-rd IEEE international conference INFOCOM*. ArXiV preprint `cs.DC/0307036`.

Jr., John M. Roberts. 2000. Correspondence analysis of two-mode network data. *Social Networks*, **22**, 65–72.

Kleinberg, J.M. 2000a. Navigation in a small world. *Nature*, **406**, 845.

Kleinberg, J.M. 2000b. The Small-World Phenomenon: An Algorithmic Perspective. *In: Proceedings of the 32nd ACM Symposium on Theory of Computing (STOC)*.

Kogut, B., & Walker, G. 2003. Restructuration ou désintégration du réseau des firmes allemandes ? *Gérer et Comprendre*, **74**.

Kogut, B., Urso, P., & Walker, G. 2006. The Emergent Properties of a New Financial Market: American Venture Capital Syndication from 1960 to 2005. *Management Science, special issue on Complex Systems Across Disciplines*.

Lind, Pedro G., González, Marta C., & Herrmann, Hans J. 2005. Cycles and clustering in bipartite networks. ArXiV preprint `cond-mat/0504241`.

Milgram, S. 1967. The Small World Problem. *Psychology today*, **1**, 61–67.

Molloy, M., & Reed, B. 1995. A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms*.

Molloy, M., & Reed, B. 1998. The size of the giant component of a random graph with a given degree sequence. *Combinatorics, Probabilities and Computation*.

Morris, Steven A., & Yen, Gary G. 2005. Construction of bipartite and unipartite weighted networks from collections of journal papers. ArXiV preprint `physics/0503061`.

Newman, M. E. J. 2004. Analysis of weighted networks. *Phys. Rev. E*, **70**.

Newman, Mark E. J. 2000. *Who is the best connected scientist? A study of scientific coauthorship networks*. E. Ben-Naim H. Frauenfelder and Z. Toroczkai (eds), Springer. ArXiV preprint `cond-mat/0011144`.

Newman, Mark E. J., Strogatz, Stevens H., & Watts, Duncan J. 2001a. Random graphs with arbitrary degree distributions and their applications. *Physics Reviews E*, **64**. ArXiV preprint `cond-mat/0007235`.

Newman, Mark E. J., Watts, Duncan J., & Strogatz, Stevens H. 2002. Random graph models of social networks. *PNAS*, **99**, 2566–2572.

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

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

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

Newman, M.E.J. 2003b. The structure and function of complex networks. *SIAM Review*, **45, 2**, 167–256.

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

Onody, Roberto N., & de Castro, Paulo A. 2004. Complex network study of Brazilian soccer players. ArXiV preprint `cond-mat/0409609`.

Peltomaki, Matti, & Alava, Mikko. 2005. Correlations in Bipartite Collaboration Networks. ArXiV preprint `physics/0508027`.

Perugini, Saverio, Goncalves, Marcos Andre, & Fox, Edward A. 2003. A Connection-Centric Survey of Recommender Systems Research. ArXiV preprint `cs.IR/0205059`.

Robins, Garry, & Alexander, Malcolm. 2004. Small Worlds among Interlocking Directors: Network Structure and Distance in Bipartite Graphs. *Computational & Mathematical Organization Theory*, **10**(1), 69–94.

Robins, G.L., Snijders, T.A.B., Wang, P., Handcock, M., & Pattison, P. 2005. Recent developments in exponential random graph (p*) models for social networks. *Social Networks*. In press.

Roth, Camille, & Bourgine, Paul. 2005. Epistemic communities: description and hierarchic categorization. *Mathematical Population Studies*, **12(2)**, 107–130.

Skvoretz, John, & Faust, Katherine. 1999. Logit Models for Affiliation Networks. *Sociological Methodology*, **29**, 253–280.

Uzzi, Brian, & Spiro, Jarrett. 2005. Collaboration and Creativity: The Small World Problem. *American Journal of Sociology*. To appear.

Véronis, J., & Ide, N. 1995. Large Neural Networks for the Resolution of Lexical Ambiguity. *Computational Lexical Semantics, Natural Language Processing Series*, 251–270.

Viger, F., & Latapy, M. 2005. Random generation of large connected simple graphs with prescribed degree distribution. *In: LNCS special issue, proceedings of COCOON'05*. To appear.

Voulgaris, S., Kermarrec, A.-M., Massoulie, L., & van Steen, M. 2004. Exploiting Semantic Proximity in Peer-to-peer Content Searching. *In: 10-th IEEE international workshop on Future Trends in Distributed Computing Systems (FTDCS)*.

Wasserman, Stanley, & Faust, Katherine. 1994. *Social Network Analysis: Methods and Applications*. Cambridge University Press. Revised, reprinted edition, 1997.

Watts, D., & Strogatz, S. 1998. Collective dynamics of small-world networks. *Nature*, **393**, 440–442.

Young-Choon, Kim. 1998. A Structural Analysis on Firm-Market Affiliation Networks in the Korean System Integration Industry. *Development and Society*, **27**(2).

Zipf, G. K. 1932. *Selective Studies and the Principle of Relative Frequency in Language*. Harvard University Press.

How to read and understand our plots.

We give in this appendix additional hints on how to read and understand the plots presented in this paper, for the readers who are not familiar with these statistical approaches. Of course, this appendix will not replace a statistics textbook, but it aims at giving sufficient intuition on the notions under concern to help the reader significantly.

Distributions.

The main statistical notion used in this paper is the one of *distribution* of a measured quantity: it is, for each possible value k of this quantity, the fraction p_k of objects which exhibit this value when the quantity is measured on them [22]. For instance, the degree distribution in a network is, for each integer k, the fraction of nodes of degree k (*i.e.* with k links).

One may consider the *number* of objects in place of the *fraction*. Both notions of distributions are strongly related, since the fraction is the number divided by the total number of objects. As a consequence, the shape of the plot is exactly the same; the only difference lies in the rescaling of the vertical axis (initially between 0 and the total number of objects, to between 0 and 1 after rescaling). Both variants have their own advantages and drawbacks. In this paper, we use the *fraction* variant to make it easier to compare between different cases: it is easier to compare the fact that in one network the fraction of degree one nodes is 0.5 (*i.e.* $50\,\%$ of the nodes have degree one) and in another one it is 0.8 (*i.e.* $80\,\%$) than the raw numbers.

In our context, the key property of the observed distributions is wether they are *homogeneous* or *heterogeneous*.

The plot of an homogeneous distribution [23] have a peak around an average value, and no object with measured value very different from this average [24]. More formally, the fraction of objects with measured value k, p_k, decreases exponentially fast when one goes away from the average value. Intuitively, this means that no object are very different from the average case concerning the observed value. This has important consequences, in particular the fact that the average is meaningful: it indicates the *normal* behavior, or what one may expect when taking an object at random.

On the contrary, some distributions are heterogeneous [25]: there are several orders of magnitude between observed values, and there is a significant number of objects for which the measured value is very different from the average one. In such cases, p_k decreases only polynomially fast when one goes away from the average value, thus much slower than in an homogeneous distribution. Then, the average value brings little information: it is not the value observed on most objects, and a randomly chosen object may exhibit a very different value. In such cases, characterising the heterogeneity of the distribution is more meaningful. This is generally done by fitting the distribution with a power-law ($p_k \sim k^{-\alpha}$ for a constant α) and then considering the

[22] *i.e.* the number of such objects divided by the total number of objects.

[23] Most famous such distributions are normal, Gaussian and Poissonian distributions.

[24] A typical example is body height: there is an average height, and nobody is twice this value high.

[25] Most famous such distributions are Zipf and power-law distributions.

exponent of this power-law (α) as a measure of the heterogenity of the distribution (lower exponents reveal higher heterogeneity, but the fact that the distribution is well fitted by a power-law is sufficient to show that it is highly heterogeneous).

Notice that it is not immediate to determine if a given distribution is well fitted by a power-law: on usual plots, the difference between exponential and polynomial decreases is not visible. This is why, when one suspects the presence of a power-law, one uses log-log scales: instead of plotting p_k as a function of k one plots $\log(p_k)$ as a function of $\log(k)$. If the distribution is a power-law, we have $p_k \sim k^{-\alpha}$, and thus $\log(p_k) \sim -\alpha \cdot \log(k)$. Therefore, the plot will be a straight line of negative slope α, which is easy to check. If the distribution has an exponential decrease, the log-log plot will not be a straight line.

On empirical data, of course, the fits are never perfect. As one may observe on the plots of this paper, however, the approach just described makes it possible to distinguish between several cases. In Figure 2, for instance, in the case of occurrence dataset, the bottom degree distribution is very well fitted by a power-law, whereas the top degree distribution certainly is not a power-law. This confirms the immediate observation that, in this case, bottom degrees span several orders of magnitudes (from 1 to more than 10000) whereas top degrees do not.

Cumulative distributions.

For several reasons, it is interesting in some situations to consider the *cumulative* distributions, instead of classical distributions as described above: one plots the fraction of objects having a measured value *lower than or equal to* k, for each k, instead of the fraction of objects having exactly this measured value.

This is particularily useful when one wants to observe the distribution of a property taking real values, not only integer ones: it is sufficient to consider a finite number of points in the plot. This is why we used cumulative distributions for our plots of clustering coefficients and redundancy (Figures 6, 7 and 10). It also helps in estimating the number of nodes with high clustering coefficients or redundancy, which is appealing in this context.

Correlations.

Finally, we present in this paper another kind of plots, aimed at observing correlations between different values attached to a same object (like the degree of a node and the average degree of its neighbors, in Figure 3). There are many way to investigate such correlations. We use here plots in which we put a dot for each object, this dot having coordinates given by the two values of interest (in Figure 3, each node leads to a dot for which x is the degree of the node and y is the average degree of its neighbors).

Such plots make it possible to observe if having a given value for one observed property is related to having a given value for another one. In particular, one may observe if having high value for the first implies a high value for the second. In the case of Figure 3, for instance, the leftmost plot of the first row (top degree correlations for the actors-movies network) shows that in random networks the average degree of neighbors of a node is independent of the degree of the node: it forms an horizontal line, indicating that it is a constant (roughly equal to 32). Instead, in

the same plot, one sees that for high degree nodes the average degree of their neighbors tends to be smaller than for lower degree nodes, thus indicating that high degree nodes are more linked to low degree nodes than others (and more than if links were random). In terms of the underlying data, it shows that if a movie has many actors, then many of these actors played in few movies only.