Multi-level analysis of an interaction network between individuals in a mailing-list

Rémi Dorat [1,2,3] Matthieu Latapy [1] Bernard Conein [2] Nicolas Auray [3]

Abstract

It is well known now that most real-world complex networks have some properties which make them very different from random networks. In the case of interactions between authors of messages in a mailing-list, however, a multi-level structure may be responsible for some of these properties. We propose here a rigorous but simple formalism to investigate this question, and we apply it to an archive of the Debian user mailing-list. This leads to the identification of some properties which may indeed be explained this way, and of some properties which need deeper analysis.

INTRODUCTION.

It makes no doubt that understanding how individuals interact in a social framework is a key issue for sociology and in many other contexts like economy, management, anthropology, etc. Collecting large-scale data on such interactions was however a challenging, almost impossible, task until recently. The birth and development of computation and communication capabilities (in particular the internet) opened unleashing opportunities for the study of interactions between individuals. Indeed, it is easy in a digital framework (as opposed to *real* world) to collect large amounts of traces of such interactions. This can be done for instance on instant messaging applications [35], at e-mail level [18], at web or blog levels [12], [1], in peer-to-peer systems [23], [31], [26], [25], and many others. The references cited here are only a few examples of the huge amount of studies conducted recently in this area, and made possible by this new situation. See [3], [43], [48], [17] for surveys of the field.

It must however be clear that the data collected this way are incomplete and often imprecise. It may be significantly biased by the measurement process, see for instance [29], [30]. Even more importantly, the behaviors of individuals themselves may be influenced by the communication medium, see for instance [7], [51]. These aspects must be taken into account in any rigorous study of individual interactions in a numerical framework.

Our contribution lies in this context. It focuses on the interaction network induced by exchanges between authors in a mailing-list. This network may be viewed as the fusion of several pieces of interactions centered on a given topic, captured by the notion of thread in the mailing-list context. One may then wonder if, and how, properties of the interaction network may be induced by this underlying structure. The aim of this paper is to answer this question.

Before entering in the core of the paper, we need some preliminaries (on the notions under study, the context and methodology, Section I). We will then present results on the analysis of the network we consider (Section II), and the multi-level formalism we propose for it (Section III). We then present results on the two indermediate levels (Sections IV and V), and we finally present and discuss our results in Section VI.

[1] LIAFA – CNRS and Université Paris 7 – 2, place Jussieu, 75005 Paris, France – `latapy@liafa.jussieu.fr` – corresponding author

[2] GSPM – CNRS and EHESS – 10, rue Monsieur le Prince, 75006 Paris, France

[3] ENST – ENST – 46, rue Barrault, 75013 Paris, France

I. Preliminaries.

It appeared recently (at the end of the 90's, [53]) that most large real-world complex networks have several properties in common. They also have specific properties which make them different from each other. These properties are useful to describe a given network (or a set of networks) and to obtain relevant information from it (them). Since a few years, many such properties have been defined and many special cases have been studied, see the surveys [3], [43], [48], [17].

Here we will observe some of these properties, starting from the most basic ones and going to more subtle ones, on an interaction network between authors in a mailing-list. We introduce these properties below. Then we describe the context in which our work lies, and the methodology we will use. Finally, we will describe precisely the raw data on which our work relies.

The observed properties.

A network is modeled by a graph $G = (V, E)$ where V is the set of nodes and $E \subseteq V \times V$ is the set of links. We will consider only undirected networks here [4], which means that we make no distinction between $(u, v) \in E$ and $(v, u) \in E$. We will denote by $N(v) = \{(u, v) \in E\}$ the *neighborhood* of a node v, the elements of $N(v)$ being the *neighbors* of v. The number of nodes in $N(v)$ is the *degree* of v: $d^o(v) = |N(v)|$.

The basic properties describing such a graph are its *size*, *i.e.* its number of nodes $n = |V|$ and its number of links $m = |E|$, its *average degree* $k = \frac{2m}{n}$, and its *density* $\delta(G) = \frac{2m}{n(n-1)}$, *i.e.* the number of existing links divided by the number of possible links. In other words, $\delta(G)$ is the probability that two (distinct) randomly chosen nodes are linked together.

Going further, one may define the *distance* $d(u, v)$ between two nodes u and v in the graph as the length of a shortest path between u and v, *i.e.* the minimal number of links one has to use to go from one node to the other. The *average distance* of the graph, $d(G)$, is nothing but the average of the distances for all pairs of nodes: $d(G) = \frac{1}{n^2}\sum_{u,v \in V} d(u, v)$. The *diameter* D of the the graph is the largest distance between any two pairs: $D = \max_{u,v \in V}(d(u, v))$.

Notice that it is possible (and in general it is true) that there are some nodes between which no path exists in the graph. To capture this, one may define the *connected components* of the graph as the largest sets of nodes such that there exists a path between any two elements of a same set. If there is only one such set, then the graph is said to be *connected*. If there are several ones, then one of them is generally much larger than the others; in such cases it is called the *giant component* and its size is denoted by $\overline{n}$.

If the graph is not connected, then there exists pairs of nodes for which the notion of distance is undefined; one then usually only considers the giant component, if it exists. We will follow this convention in this paper. Therefore, we consider that the notions of distance defined above only concern the giant component (we will see that the networks we will encounter all have a giant component).

The next property is not this classical. It is the *degree distribution*, *i.e.* for all integer i the number of nodes of degree i. One may also observe the correlations between degrees, defined as the average degree of the neighbors of nodes of degree i, for each integer i. Other notions concerning degrees have been studied, like assortativity, but we do not use them here.

Another important kind of statistics aims at capturing a notion of local density: it measures the probability that two nodes are linked together, provided they have a neighbor in common. In other words, it is the probability that any two neighbors of any node are linked together. This is measured using the *clustering coefficient* of a node v:

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

[4] We will discuss this choice later in the paper.

where $E_{N(v)} = E \cap (N(v) \times N(v))$ is the set of links between neighbors of v. In other words, $cc(v)$ is the probability that any two neighbors of v are linked together. Notice that it is nothing but the density of the neighborhood of v, and in this sense it captures the local density. It is undefined for nodes of degree lower than 2.

The clustering coefficient of the graph itself then is the average of this value for all the nodes on which it makes sense: $cc(G) = \frac{1}{|\{v,\ d^o(v)>1\}|} \sum_{v \in V,\ d^o(v)>1} cc(v)$. Other notions of clustering coefficients have been defined to capture local density but this one is sufficient for our purpose.

The distance may also be used to define a notion of *centrality* of nodes [52]. Let us denote by $d(v)$ the average distance of v to any node in the graph: $d(v) = \frac{1}{n} \sum_{u \in V} d(v,u)$. Then one may consider that v is more *central* than u when $d(v)$ is smaller than $d(u)$. Other notions of centrality (like the degree itself or the *betweenness centralty* [52]) are often used, but they are out of the scope of this paper.

All these notions naturally lead to the observation of their distributions and of their possible correlations, which we detail now.

The distribution of an integer valued property is, for all integer i the number of instances (nodes or pairs of nodes in our context) for which this property has value i. For instance, the degree distribution is the number of nodes having degree i, for all i. If the property is real-valued (like the clustering for instance), we take for all integer i the number of instances for which the property has a value between $\frac{i-0.5}{100}$ and $\frac{i+0.5}{100}$ and we plot it as a function of $\frac{i}{100}$. Distributions make it possible to observe the representativity of the average value, and to identify non-typical cases.

Correlations between a property P and another property P' are usually captured by plotting the average value of property P' for nodes for which property P has value i, for all i. For instance, the degree-degree correlations are studied by plotting, for each i, the average degree of neighbors of nodes of degree i. The degree-clustering (resp. degree-centrality) correlations are studied by plotting the average clustering coefficient (resp. centrality) of nodes of degree i as a function of i. The clustering-centrality correlations are studied by plotting the average centrality of nodes of clustering between $\frac{i-0.5}{100}$ and $\frac{i+0.5}{100}$ as a function of $\frac{i}{100}$. These plots make it easy to observe how a property tends to be related to another one, for instance if highest degree nodes tend to be linked to highest degree nodes or not, if they tend to have a high clustering or not, and/or if they tend to have a high centrality or not.

One may of course consider many other statistics to describe complex networks. We will focus here on the statistics described above, which play a central role in complex network studies and already provide a powerful toolkit for their analysis.

Typical complex networks.

It appeared recently [53] that most large real-world complex networks have several non-trivial properties in common. First notice that, since we are concerned here with large networks, n must be large. In most real-world cases, is appeared that m is of the same order of magnitude as n, *i.e.* the average degree is small compared to n. Therefore, the density generally is very small: $\delta(G) \sim \frac{2kn}{n(n-1)} \sim \frac{1}{n}$, which is close to 0 since n is large.

It is now a well known fact that the average distance and the diameter in real-world complex networks are in general very small (*small-world* effect), even in very large ones, see for instance [36], [53]. This is actually true in most graphs, since a small amount of randomness is sufficient to ensure this, see for instance [53], [33], [19], [9], [42]. This property, despite it may have important consequences and should be taken into account, therefore should not be considered as a significant property of a given network. We will discuss this in the methodology part below.

Another point which recieved recently much attention, see for instance [21], [5], [4], is the fact that the degree distribution of most real-world complex networks is highly heterogeneous, often well fitted by a power law: $p_k \sim k^{-\alpha}$ for an exponent α generally between 1 and 3.5. This means that, despite most nodes have a (very) 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 graph with the same size (*i.e.* same number of nodes n and links m) as a given real-world one [5], thus with the same density, then the obtained degree distribution is qualitatively different: it follows a Poisson law (in which all the values are close to the average). This means that the heterogeneous degree distribution is not a trivial property, in the sense that it makes real-world complex networks very different from most graphs (of which random graphs are representative). The degree correlations and other properties on degrees, on the countrary, behave differently depending on the complex network under concern.

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

If, like above, one samples a random graph with the same size as a real-world complex network then it clustering coefficient is equal to the density. It is therefore very low. If one samples a random graph with the same number of nodes *and* the very same degree distribution [5] then the clustering coefficient still is significantly smaller [43] than in real-world cases. The clustering coefficient therefore captures a property of networks which is not a trivial consequence of the degree distribution.

Finally, the vast majority of large real-world complex networks have a very low density, small average distance and diameter, a highly heterogeneous degree distribution and a high clustering coefficient. These two last properties make them different from random graphs of the same size (both purely random and random with prescribed degree distributions). As we will see in Section II, this is also true for the network we consider here. More subtle properties may be studied, but until now no other one appeared to be a general feature of most real-world complex networks. The properties described here therefore serve as a basis for the analysis of real-world complex networks, with additional properties used to describe special cases of interest.

Context

Many real-world complex networks have been studied using the properties described above. Let us cite for instance file sharing [32], [23], [31], [50], [26], [25], company boards [46], [14], [6], [42], sport teams [10], [44], movie actors [53], [42], human sexual relations [20], [34], attendance to political events [22], financial networks [13], [15], [24], [54], recommandation networks [45], theatre performances [2], [49], politic ativism [11], and scientific authoring [47], [39], [41].

Since, as explained above, some of their properties appear to be very general, much effort has been done in searching for underlying principles to explain them. The most famous one probably is the preferential attachement principle [5]. Nodes arrive one by one and are linked with pre-existing nodes with a probability proportional to their degree. The idea is that individuals tend to link themselves to popular persons, thus increasing their popularity. This induces power law degree distribution, and this principle is nowadays the most widely accepted explanation for this property. Other attempts have been done for various properties, see for instance [40], [17], [27].

When one turns to more precise properties, like the exact degree distribution, the clustering coefficient, or more subtle properties, it is however difficult to explain them as consequences of simple principles. One then often refers to complex notions related to the semantics of the links and nodes, to possible behaviors of individuals (like the fact that they tend to introduce each other to their friends or more complex principles), etc. These assumptions are difficult to validate (measuring them is a challenge in itself), which makes it hard to evaluate these efforts and their results.

[5] We consider here graphs chosen uniformly at random in the set of all graphs having the prescribed properties, using typically the Erdös and Rényí or the *configuration* models [19], [9], [8], [37], [38].

We use here quite a different approach. We try to explain the observed properties of the network we consider (both simple and more subtle ones) as consequences of its multi-level nature: it is constructed by merging many small networks derived from the threads. These small networks, and the merging process itself, have their own properties, which may be responsible for many properties of the global network. In other words, we seek *structural*, as opposed to semantic, explanations of these properties.

Interestingly, one may have a different view regarding our contribution. One may notice that semantic features actually are encoded in the multi-layer construction of the network. For instance, one may imagine that the topic of the exchanges in the mailing lists and the author behaviors are somewhat encoded in the thread structure and in the construction process. This is certainly true, and our contribution may therefore also be viewed as a way to investigate how much of these semantic aspects is encoded in the thread structure and their combination.

Methodology.

As sketched above, the main methodology developped in recent years for the analysis of real-world complex networks relies on the definition of properties describing these networks and on comparison of real-world networks with random graphs. The underlying idea is that a property makes sense if it is *not* typical of *all* networks having the other properties, *i.e.* networks choosen uniformly at random among these networks.

According to this approach, for instance, the low average distance met in practice is not a significant property, as it also is a property of any network with a reasonable amount of randomness, including random networks and random networks with prescribed degree distribution. Instead, the heterogeneous degree distribution is significant since it is in sharp contrast with what is met in random networks. If one takes a random network among the ones having this degree distribution, then the clustering coefficient remains low, which leads to the conclusion that this property also is significant: it is not present in most networks, and is not a trivial consequence of the degree distribution. One can push further this approach with any property of a network, and with any model aimed at capturing some of these properties. The properties met in practice which are not fitted by models reveal a real-world feature which is not captured by the model, and so it is significant.

We will use this methodology in this paper. We will compare the objects under study to comparable random structures, and we will propose simple models to capture the observed properties. We will focus on the way the network we consider is constructed, and we will mimic this construction process from random structures in order to see if the properties of the obtained network are comparable to the ones of the original network. If this is the case, we will conclude that these properties may be seen as consequences of properties of the construction process. We will seek both properties which fit in this framework and properties which do not, in order to make the difference between somewhat trivial properties and properties which need more investigation.

Notice finally that the random structures may be formally studied. This however often is very hard and leads to approximate results which may not fit the reality very well. Instead, one may generate many random objects in the considered class and then take the average behavior. This is what we will do here.

The data.

Our contribution, despite it can be seen as very general, relies on the use of a real-world usage trace. It is a set of messages posted on the *Debian* mailing-list, the archives of which being available online [16]. The selected data corresponds to exchanges processed during one year, from august 2003 to august 2004, on the French mailing-list.

The data contains $25\,941$ messages posted from $2\,287$ different e-mail addresses, corresponding to $6\,731$ threads. We will consider that each e-mail address corresponds exactly to one individual, which is not true in practice (both indviduals may have several addresses, and an address may be used by several individuals). This however has little influence, if any, on the results we derive here.

Let us insist on the fact that this dataset is considered here as an example of the kind of data to which our approach may be applied. In particular, we consider it as representative of exchanges in a mailing-list, despite its particular nature (the fact that it is a newgroup, its technical content, etc) may have an impact on its properties. Indeed, we will focus on very general properties of exchanges in mailing-lists, and we will not derive results on particular aspects of this data. We will discuss this further in Section VI.

II. The interaction network.

The central object in this paper is the interaction network between authors of the e-mails in the database described above. Some of these e-mail are answers to others, and this induces a relation between them, which can be transposed to authors: if there is in the data an e-mail authored by u which is an answer to an e-mail authored by v, then we say that u answered to v.

We then model the interaction network as the graph $G = (V, E)$ where V is the set of all the authors (identified by an e-mail address as explained above) and where $(u, v) \in E$ means that u answered to v, or v answered to u.

In this paper, we consider G as undirected: no distinction will be made between (u, v) and (v, u). In other words, $(u, v) \in E$ implies that u answered to v, or v answered to u, or both. We also remove loops, *i.e.* links of the form (v, v). These simplifications induce some loss of information but it is not crucial in our context where we want to study global statistics on the network. Instead, it helps much in simplifying the involved notions since most studies until now considered undirected loop-free networks (and so the properties are defined on such graphs).

Likewise, one may consider a weighted graph by adding on each (directed or not) link (u, v) the number of times u answered to v in the dataset. Again, this would encode much more information than the unweighted graph we consider, but it would make its analysis much more intricate. Moreover, there in no need of this additional information for our purpose. We will discuss this further in Section VI.

We can now observe the various properties of this network. The most basic ones are shown in Table I. The degree distribution and degree correlations are given in Figure 1. The clustering coefficient distribution and its correlations with degrees are given in Figure 2. The distribution of connected component sizes and the distribution of distances between pairs of nodes are given in Figure 3. The distribution of centrality is very similar to the one of distances between pairs therefore we do not present it here. Instead, we display in Figure 4 the correlations of both degree and clustering with the centrality. In all the relevant cases, we also give the values and display the plots obtained for random graphs with the same size and for random graphs with the same size and the same degree distribution.

	nb nodes	nb links	avg degree	density	component	avg distance	diameter	clustering
	n	m	k	δ	$\overline{n}$	d	D	*cc*
original	2287	9592	8.39	0.0037	1743	2.97	8	0.33
purely random	–	–	–	–	2285	3.87	7	0.0042
random with degrees	–	–	–	–	1751	2.90	7	0.29

Table I. Basic statistics for the interaction network.

The first point here is to observe that our network has all the properties typical of real-world complex networks. Its average degree is low compared to its number of nodes, thus its density is very small. Its degree distribution is very heterogeneous, with more than $50\,\%$ of nodes having less than 5 links (536 have no link at all), but some nodes with degree around 400. This means that some authors received no answer (the ones with degree 0) while others interacted with a significant portion of all the authors. The clustering coefficient itself is large compared to the density: two nodes are linked together with a probability approximately 100 times higher if they have a neighbor in common than if they are chosen at random. The network has a giant connected component and both its average distance and its diameter are quite small, as expected.

Going further, we may observe that the average degree of the neighbors of a node is significantly related to its own degree. Small degree nodes tend to be connected to high degree ones, and conversely. Likewise,

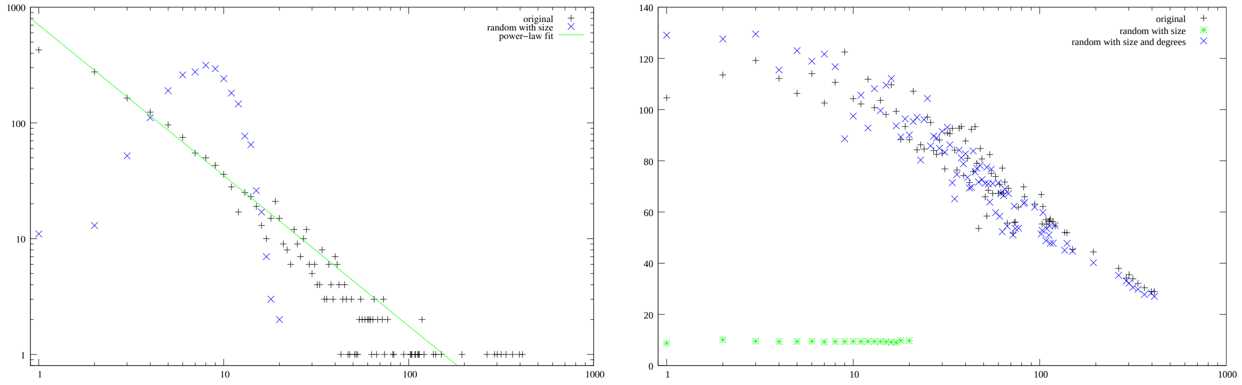

Fig. 1. Left: the degree distribution of the original interaction network, fitted by a power law of exponent $\alpha = 1.3$, and the one of a typical random graph of same size. Right: the degree correlations, *i.e.* the average degree of neighbors of nodes of degree i as a function of i, for both the original interaction network, for a typical random graph of same size, and for a typical random graph with the same size and degree distribution.

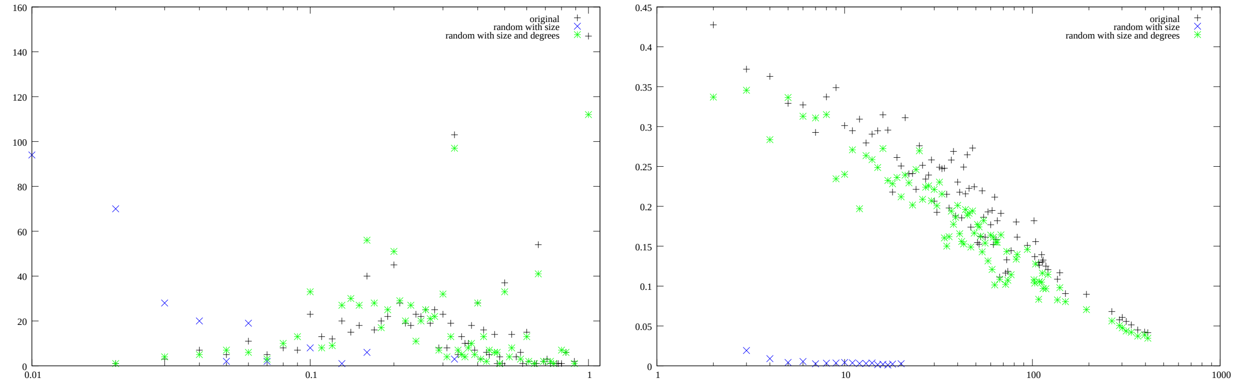

Fig. 2. Left: the clustering coefficient distribution. Right: the correlations between clustering coefficient and degree, *i.e.* the average clustering coefficient of nodes of degree k as a function of k. Each plot is given for both the original interaction network, for a typical random graph of same size, and for a typical random graph with the same size and degree distribution.

small degree nodes tend to have a high clustering while high degree ones have a smaller clustering. The network has many nodes of degree 0, which induces the same number of connected components of size 1. It also has 8 components reduced to only one link, and all the other nodes are in the giant component. It may therefore be viewed as connected, once the nodes of degree 0 have been removed. In the giant component, the distances are well centered on an average value: only a few pairs of nodes are at a distance which varies significantly from the average, and even in these cases the difference remains small. Finally, it appears clearly in Figure 4 that nodes with high degree are more *central* in terms of distance than nodes with low degree. On the countrary, there is no obvious relation between clustering and centrality.

Let us insist on the fact that our purpose here is *not* to interpret these results: our aim in this section was to identify some non-trivial properties of the network under concern, in order to explore in the next sections how the way it is constructed may be seen as responsible for these properties.

It appears clearly that the interaction network is very different from a random graph with the same size: the degree distribution is heterogeneous, the clustering coefficient is several orders of magnitude larger than in a random graph, and actually all the other properties are poorly fitted by random graphs, see the figures. Notice that the fact that there are very few nodes of very low degree in purely random graphs implies that it is almost connected (the giant component is almost the whole graph). If we first remove all the nodes of degree 0 or if we restrict ourselves to the giant component of the original network, however, the results are similar: the original interaction network is far from a random graph of the same size.

If we compare it to a random network with the same size and degree distribution, the difference is not so huge. First, of course, the degree distribution is the same, which implies that there is the same

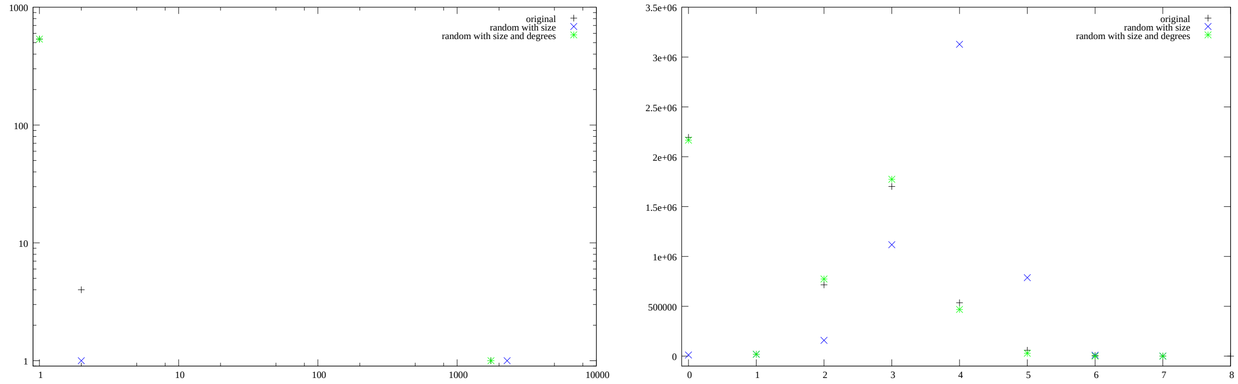

Fig. 3. Left: the connected component size distribution. Right: the distribution of distances between pairs of nodes. Each plot is given for both the original interaction network, for a typical random graph of same size, and for a typical random graph with the same size and degree distribution.

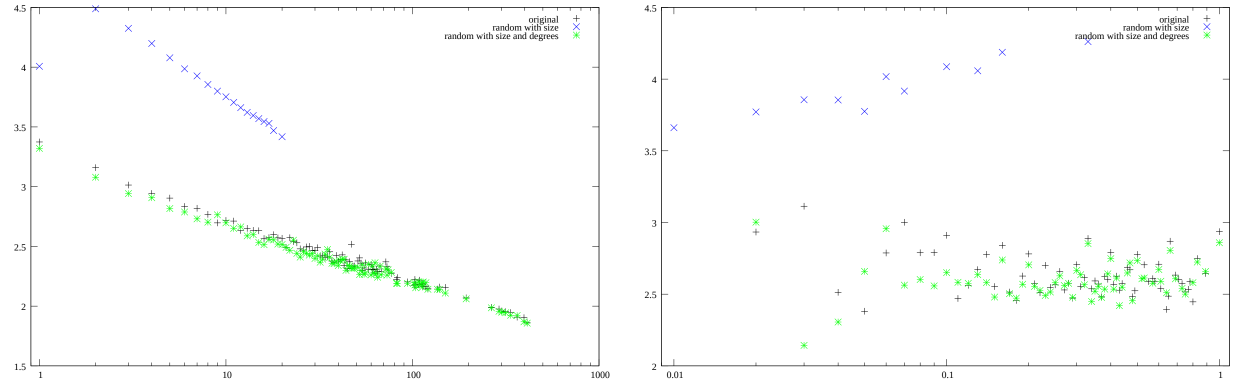

Fig. 4. Left: the correlations between centrality and degree, *i.e.* for all i the average distance of nodes of degree i to all others. Right: the correlations between centrality and clustering, *i.e.* for all $\frac{i}{100}$ the average distance of nodes of clustering between $\frac{i-0.5}{100}$ and $\frac{i+0.5}{100}$ to all others. Each plot is given for both the original interaction network, for a typical random graph of same size, and for a typical random graph with the same size and degree distribution.

amount of nodes of degree 0, almost all the others being in the giant component. Therefore the size of the giant component and the distribution of the connected component sizes are well fitted. Likewise, the degree correlations and the distance distribution are very well fitted, which means that they may be seen as consequences of the size and the degree distribution. The fit for correlations between degree and centrality is also quite good.

Though the difference is not huge, the fit is not as good if we observe properties related to clustering. First, the average clustering is significantly lower in random graphs with the same size and degree distribution than in the original network. As can be observed in Figure 2, the clustering distributions have the same shape but the original one is shifted towards the largest values. The correlations with degree show that this is due to the fact that nodes of low degree (in particular the ones with very low degree) tend to have a very high clustering in the original network: almost $50\,\%$ of nodes of degree 2 actually form a triangle with their two neighbors (while only one third do in the corresponding random graph).

Finally, we obtain quite a precise description of the interaction network we consider (though many other properties may be observed), and we give evidence of the fact that is is very different from a typical random graph with the same size. The fit with a random graph with the same size and degree distribution is much better, but not perfect. Moreover, obtaining properties as a consequence of global statistics like the degree distribution is not satisfactory since it brings unsufficient explaination of the *causes* of these properties. Moreover, as we will discuss in Section VI, this approach can hardly be extended to more subtle properties. This is why we propose another approach aimed at capturing the original properties more precisely, at giving some explainations for these properties, and which may be extended to more complex properties.

III. The multi-level formalism.

The raw data is nothing but a set of messages, which we will denote by M. Each message m is labelled with an author $a(m)$. Moreover, m may be an answer to another message m'. We then call m' the *father* of m and we denote it by $m' = f(m)$. If m has no father defined this way (it is not an answer to any other message) then we put as a convention that $f(m) = m$.

This leads to the following set of definitions. The *root* $r(m)$ of a message m is either m itself if $f(m) = m$, or else it is the root of $f(m)$. Notice that not all message is the root of any message, but only the ones which are not answers to any other message. Given the nature of our data, we call these messages the *roots*, or *queries* (they generally correspond to queries posted by users on the mailing-list) and denote their set by $Q \subseteq M$.

We may now define the *thread* to which a message m belongs as

$$t(m) = \{m' \text{ such that } r(m') = r(m)\}$$

A thread t then is a set of messages such that all messages in the set have the same root and no other does. We will denote the set of threads by T. Notice that a thread t always contains exactly one root, which we denote by $r(t)$, and each root r defines exactly one thread, $t(r)$. Therefore there is a trivial bijection between the set of threads, the set of roots and the set of queries. We will use these terms equivalently, depending on the context.

A thread has a tree structure with respect to f, which leads to the following definitions. First notice that the root of a thread t is nothing but the root of the corresponding tree. Then we define the depth of a message as its distance to its root: *depth*(m) is 0 if m is a root, and $1 +$ *depth*$(f(m))$ else. The height of a thread t is the maximal depth over all its messages: *height*$(t) = max\{$*depth*$(m),\ m \in t\}$. The degree $d^o(m)$ of a message m is the number of messages $m' \neq m$ such that $f(m') = m$.

Considering now the author point of view, we define the *contribution* of an author x as the number of messages he authored: $c(x) = |\{m \in M,\ a(m) = x\}|$. Likewise, the *dispersion* of an author x is the number of threads to which he/she contributed: $d(x) = |\{t \in T,\ \exists m \in t,\ a(m) = x\}|$. Conversely, the number $a(t)$ of authors in a thread t is $a(t) = |\{a(m),\ m \in t\}|$.

The first level at which we will consider the data is this one: we see the data as a set of threads, themselves viewed as trees.

The second level at which we will consider the data is obtained from the first one by adding the authoring information: each thread is a labelled tree.

Finally, the third level is the one of the interaction network, already defined and studied in Section II. It can be defined using the formalism above as follows: $G = (V, E)$ where $V = \{a(m),\ m \in M\}$ is the set of authors, and $E = \{(u, v),\ u = a(m) \in V,\ v = a(m') \in V,\ m \neq m',\ m = f(m') \text{ or } m' = f(m)\}$ is the set of links such that two authors are linked if one of them answered to a message posted by the other. Notice that this graph may be obtained from the thread tree structures by merging all the nodes having the same author.

The three levels are illustrated in Figure 5. It must be clear that the data may be considered at several other levels, and could be observed using a variety of models. For instance, one may consider the threads as graphs among authors. One may also include the directed nature of links, or time information (the date at which each message appeared), which is available. All these formalisms may be relevant depending on the aim of each study. We focus here on the three levels defined above, which are sufficient for our purpose.

IV. The threads.

In this section we present basic statistics and models for the data at thread level. We will therefore consider sets of trees which we describe using statistical tools, and we compare the values obtained for

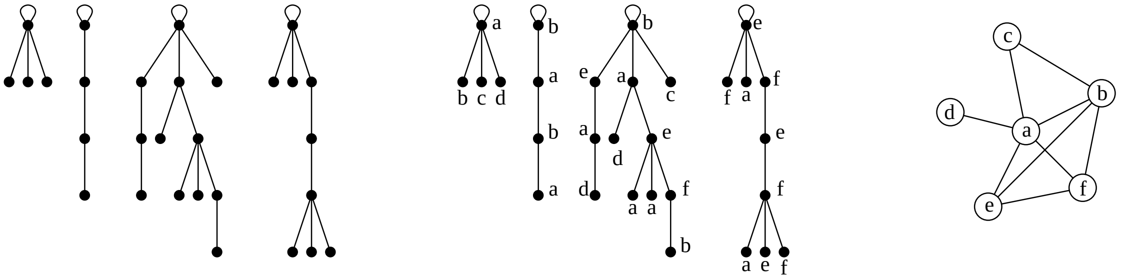

Fig. 5. The three levels at which we will consider our data. From left to right: the threads (trees), the labelled threads, and the interaction network. Notice that we removed the loops (here, (f,f)) and that we do not consider multiple links (for instance here (a,b)).

the original data to the ones obtained for the models.

The first model is the purely random one: we consider the same number of messages as in the original data, we choose randomly as many roots as in the original data, and each message is linked to a randomly chosen father. We repeat this until there is no cycle, and therefore we obtain a set of trees chosen at random among the ones having the same number of messages and roots. We will call this the *random* model for threads.

The other model we will consider only adds the degree constraint: we draw the degree of each message according to the original degree distribution and then we choose for each message a father which still has not as many sons as its degree. Again, we repeat this until there is no cycle, and therefore we obtain a set of trees chosen at random among the ones having the same number of messages and roots, and the same degree distribution as the original one. We call this model the *degree* model.

As we will see, this model is sufficient to capture the basic properties we will consider here. Moreover, it is important for us to consider only very simple models, in order to focus on the multi-level nature of the data. We will therefore not consider more subtle models.

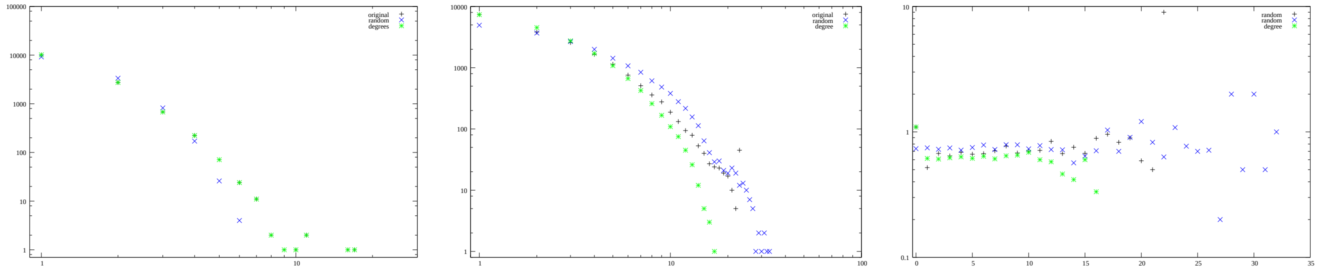

Fig. 6. From left to right: the degree distribution of messages in threads; their depth distribution; the correlations between their depth and degree. Each plot is given for the original data and both *random* and *degree* models.

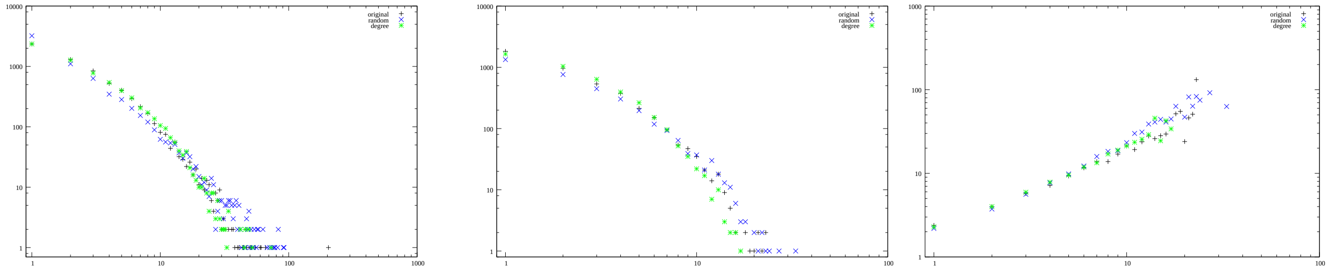

Fig. 7. From left to right: the distribution of thread sizes (number of messages); the distribution of their heights; and the correlations between both (*i.e.* the average size of threads of height i, for all i). Each plot is given for the original data and both *random* and *degree* models.

We can now observe the statistics obtained for real-world data, together with the statistics obtained for the two models. Let us begin with some properties of the messages, namely their degree distribution, their depth distribution, and the correlations between these two properties, see Figure 6. One can observe on these plots that the properties are quite heterogeneous and that there is no clear correlations between them. For instance, almost $10\,000$ messages recieved only one answer, while some recieved more than 10. There is however no message with a huge number of answers, which is not surprising. Similar remarks hold for depths.

If we turn to properties of threads themselves, the heterogeneity is more pronounced, see Figure 7: most threads contain only a few messages, but one of them contains more than 200 messages. It is however a very special case, and here again the heterogeneity is not huge. As one may expect, there is a correlation between thread height and size.

We observed various other statistics (including the correlations between the ones plotted here) and all the results are similar. We finally conclude that the *degree* model preforms better than the *random* one but the difference is not huge (which is due to a quite low heterogeneity), and the fit is good but not perfect.

It must however be clear that these models miss important properties of the threads, like for instance the presence of large filiform structures, *i.e.* series of messages m_0, m_1, …, m_l such that $m_i = f(m_{i-1)})$ and $d^o(m_i) = 1$ for all $0 < i \leq l$. Capturing such properties can be done quite easily, but it is out of the scope of this paper, see Section VI.

V. Authors in threads.

The thread models proposed in the previous section are not sufficient for our purpose. Indeed, in order to be able to construct an artificial interaction network between authors from a set of artificial threads, we need to associate an author to each message. This is the aim of this section.

Again, we will observe basic properties of this association in our real-world data, and try to capture them in very simple models. Let us suppose that a set of messages M is given and that there is a thread structure on this set defined by the function $f(m)$ which, for each $m \in M$ gives its father. We also suppose that a set A of authors is given. We want to define models which produce functions from M to A giving an author $a(m)$ to each message m.

Again, the first model we will consider is purely random: the author of each message in M is chosen uniformly at random in A. We will call this the *random* model for authors.

The other model we will consider relies on the distribution of author contributions. We suppose that this distribution is given, then we sample the contribution $c(a)$ of each author a according to this distribution, and we choose at random $c(a)$ messages $m \in M$ for which we put $a(m) = a$. We will call this the *contribution* model.

Let us notice that we may use artificial threads obtained in previous section to evaluate our models of author labelling. However, this would imply that the performances we observe in this section could be biased by the models in the previous section. We will therefore use here the original threads, and simply replace the original authors with authors chosen with the models. This makes it possible to evaluate the properties of the two kinds of models separately.

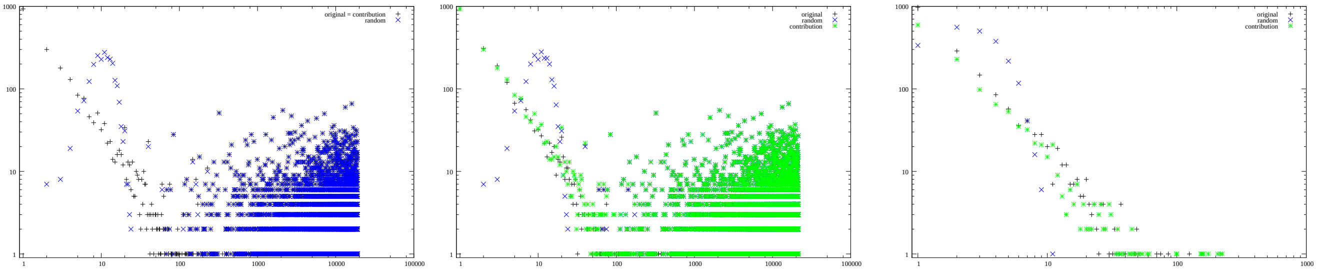

Fig. 8. From left to right: the contribution distribution; the dispersion distribution; the distribution of the number of roots labelled by the same author. Each plot is given for the original data and both *random* and *contribution* labelling models, on the original threads.

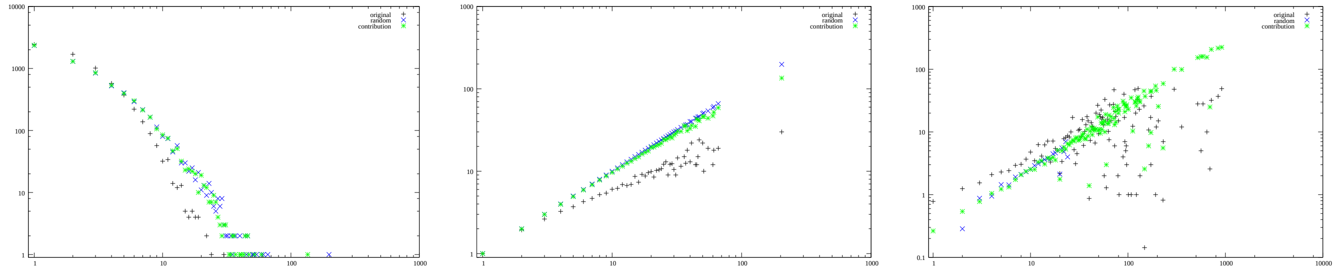

Fig. 9. From left to right: the size (in terms of authors) distribution of threads; correlations between thread sizes in terms of messages and in terms of authors; the number of roots of threads authored by each author, as a function of the total number of threads he/she authored (for each author we draw a point with coordinates given by these two properties). Each plot is given for the original data and both *random* and *contribution* labelling models, on the original threads.

Let us first observe in Figure 8 the contribution and the dispersion distributions. The shapes of the plots for the original data are unusual: they begin with a polynomial decay but the tail of the distribution in unstructured. This means that authors may be separated into two sets: the ones which have a quite low contribution, the number of which decays polynomially with the contribution, and the ones with high contribution, between which there is no difference. In other words, the number of authors having a given contribution is independent of this contribution when it is large enough. The same observations hold for dispersion. Notice that the polynomial decay is not captured by the *random* model, but that the tail is well fitted which indicates that it is due to the structure of threads rather than the labelling model. The *contribution* model takes the contribution distribution as a parameter, but it also fits the dispersion distribution very well. This is also true for the number of roots labelled by each author. We do not enter in more details here since our aim is not to give interpretations of the observed properties.

If we turn to more complex properties, like the ones in Figure 9, the fit is not so good but it remains reasonable. This shows that, as long as we are concerned with basic properties of authors in threads, the *contribution* model is sufficient. It must be clear however that it misses some important features of the original data. For instance, in the original data, if a message is authored by a then many other messages in the thread $t(a)$ containing a will also be authored by a with high probability. These properties may be included in author models, but this is out of the scope of this paper. Our purpose here is not to model the original data as precisely as possible, but to capture some nontrivial properties which may play a role in the properties of the interaction network. We will therefore not deepen more the modeling of message labels.

VI. Results and discussion.

In Section II, we described the main properties of the interaction network, up to a quite high level of detail. In Section III, we proposed a formalism which makes it natural to observe the object under concern at three different levels: the thread level, the labelled thread level, and the interaction network itself. We studied basic properties of the two first levels in Sections IV and V, and we proposed simple models to capture them.

We can now address the central question of this paper: can the properties of the interaction network be seen as consequences of properties at the two other levels? In order to answer this question, we will generate articifial networks using the models proposed for the first levels and compare them with the original network. We obtain seven artificial networks, plus the classical comparison with purely random graphs and with random graphs with the same degree distribution already considered in Section II.

We therefore produce here the same statistics as in Section II for the seven new relevant cases. See Table II and Figures 10 to 17.

There are several important points to notice. First, it appears clearly that the model used for the threads has little influence on these results. This is a consequence of the fact that, at least concerning the properties under concern, the properties of threads are quite close from random as seen in Section IV. On the

		THREADS								
		random			degree			original		
		m	k	δ	m	k	δ	m	k	δ
LABELS	random	19111	16.71	0.0073	19129	16.73	0.0073	19119	16.72	0.0073
	contribution	14415	12.61	0.0055	14450	12.64	0.0055	14420	12.61	0.0055
	original	–			–			9592	8.39	0.0037

		THREADS					
		random		degree		original	
		d	D	d	D	d	D
LABELS	random	3.01	5	3.01	5	3.01	5
	contribution	2.89	7	2.88	7	2.87	6
	original	–		–		2.97	8

		THREADS					
		random		degree		original	
		$\overline{n}$	cc	$\overline{n}$	cc	$\overline{n}$	cc
LABELS	random	2287	0.0082	2286	0.0082	2287	0.0086
	contribution	2149	0.32	2178	0.33	2192	0.33
	original	–		–		1743	0.33

Table II. Properties of the artificial interaction networks. From top to bottom: the basic statistics; the average distance and the diameter; the size of the giant component and the clustering coefficient.

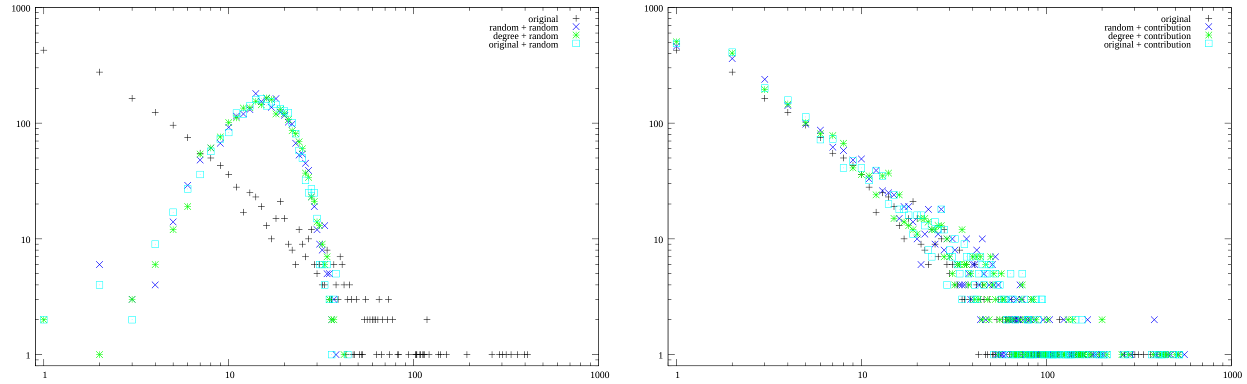

Fig. 10. Degree distributions in the artificial interaction networks. See Figure 1 and its caption.

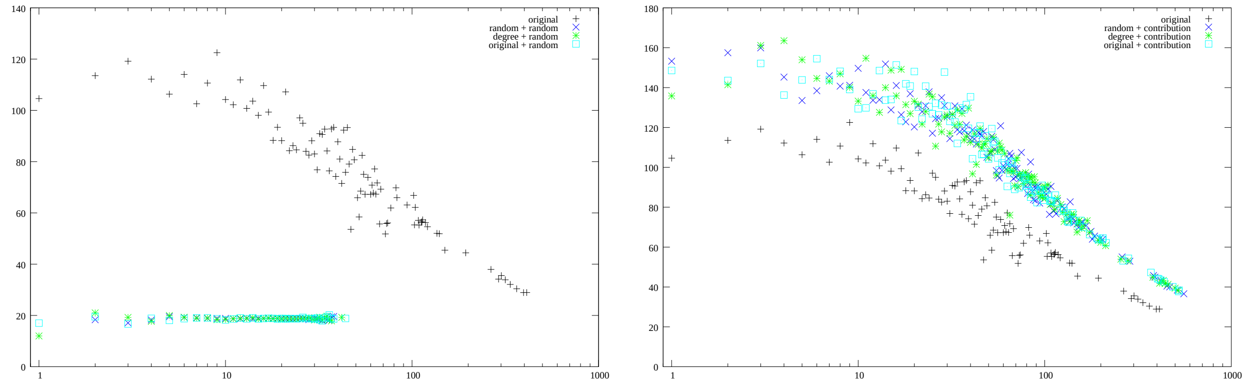

Fig. 11. Degree correlations in the artificial interaction networks. See Figure 1 and its caption.

countrary, the model used for author labellings has a strong influence, and the *contribution* model gives

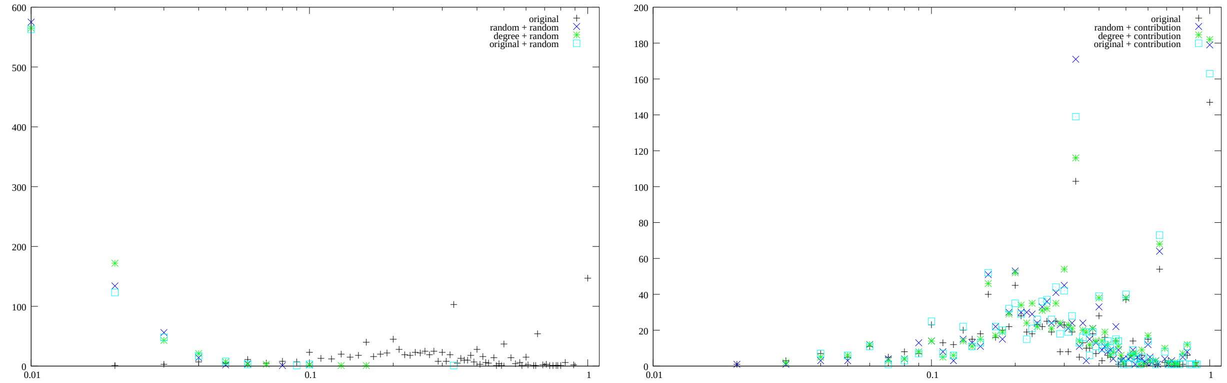

Fig. 12. Clustering distributions in the artificial interaction networks. See Figure 2 and its caption.

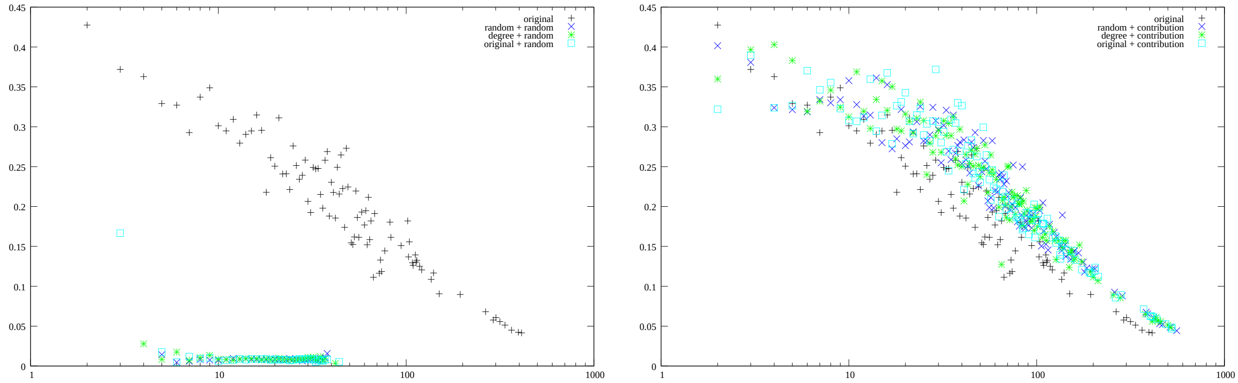

Fig. 13. Correlations between degree and clustering in the artificial interaction networks. See Figure 2 and its caption.

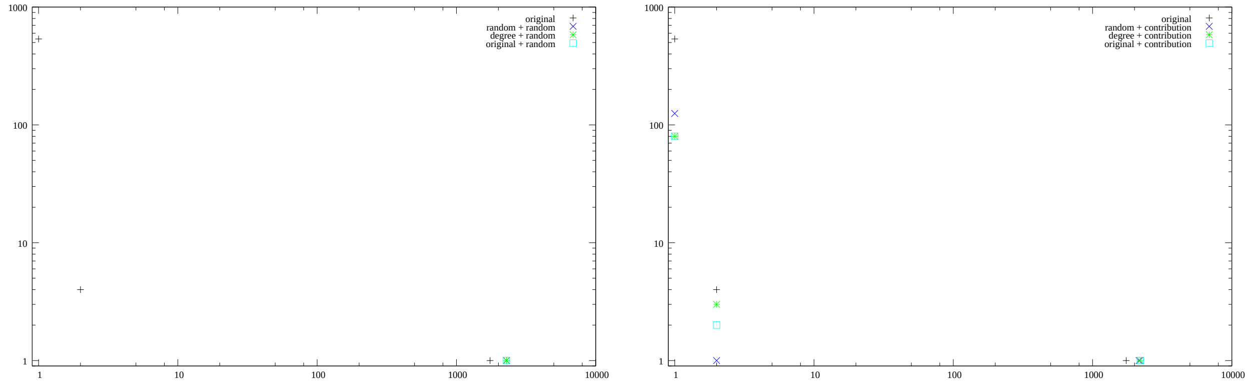

Fig. 14. Distribution of connected component size in the artificial interaction networks. See Figure 3 and its caption.

very good results. The artificial interaction networks obtained with the *contribution* model for authors and the *degree* one for threads gives better performance than the ones obtained in Section II. The only properties on which they perform poorly is the size of the giant component and the degree correlations; this is due to the fact that the artificial networks are almost connected, which is in turn due to the fact that they do not capture the presence of threads of size 1. This can be easily added in the models, or one may study these special threads separately.

Finally, it appears from these statistics that, despite our models are very basic (and, as we have seen, they miss important properties of the original data), they are sufficient to capture most simple properties of the original interaction network. In particular, they do significantly better than random graphs with the same size, and random graphs with the same size and degree distribution.

We will not go further in the analysis of the results since this is sufficient for our purpose. But we want

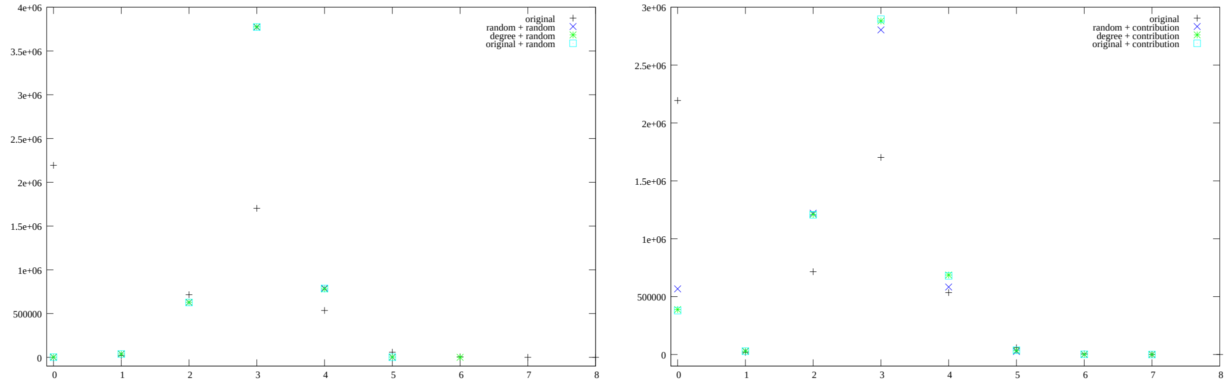

Fig. 15. Distribution of distances in the artificial interaction networks. See Figure 3 and its caption.

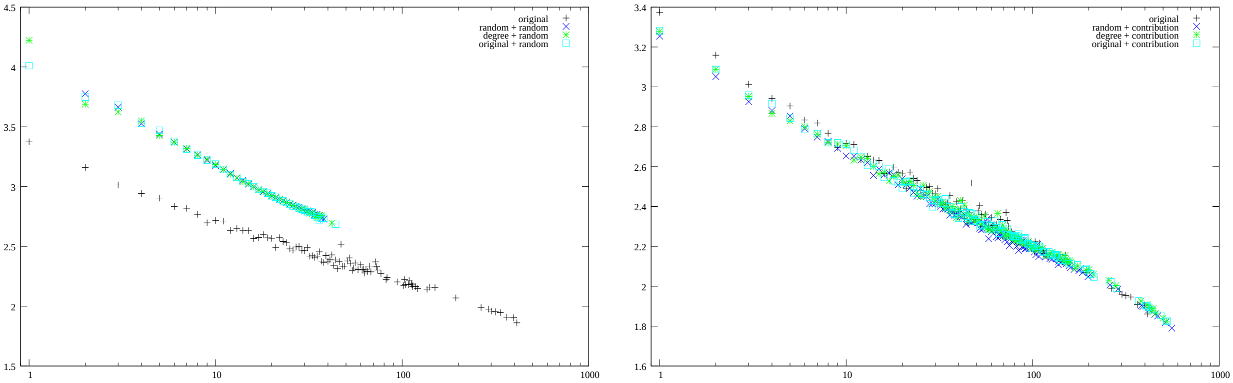

Fig. 16. Correlations between degrees and centrality in the artificial interaction networks. See Figure 4 and its caption.

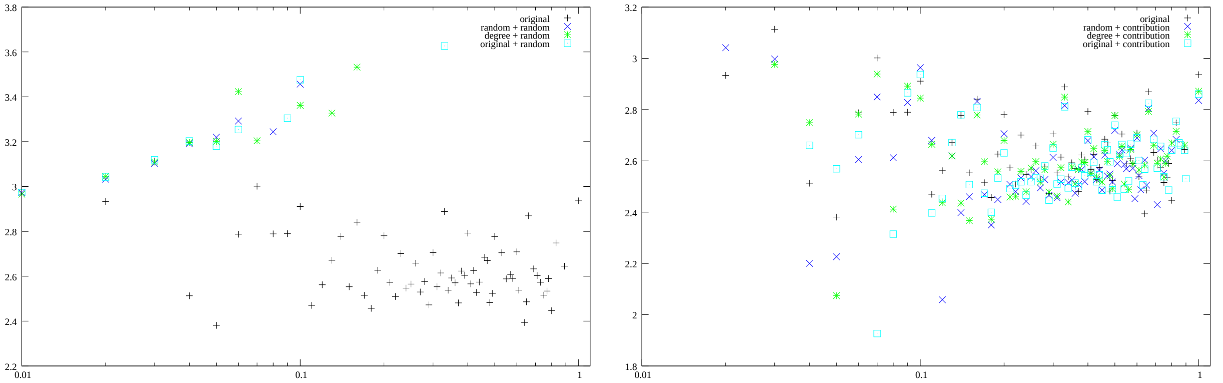

Fig. 17. Correlations between clustering and centrality in the artificial interaction networks. See Figure 4 and its caption.

to insist on one point which seems particularily important to us. It must be clear that the fact that the properties of our artificial networks are similar to the ones of the original network is a non-trivial result: these properties were not encoded explicitely in the models, which rely only on very basic properties of threads and authors. Showing that the properties of threads have little influence while the frequency of occurences of authors are central also is a non-trivial result. The multi-level formalism makes it possible to derive such results, which improve significantly our understanding of the underlying object, whereas random graph models can only be used to mimic the properties of the object.

Going further, we beleive that a multi-level approach would make it possible to capture much more subtle properties than the ones discussed here. For instance, the redundancy of authors inside each thread may induce clusters in the interaction network; the presence of filiform structures may induce large cycles; etc. More importantly, if one wants to capture the directed and/or weighted nature of the data, then the multi-level approach seems very well suited whereas random graph approaches are of limited help.

CONCLUSION AND PERSPECTIVES.

In this contribution we studied an interaction network between authors induced by exchanges in a mailing-list. We proposed a three-level formalism to describe and study this data. This formalism emphasizes the fact that the final network is constructed from smaller, simplier substructures (the threads and the labelled threads). It makes it possible to investigate the influence of the properties of these small structures, and of this construction process, on the properties of the overall network.

We observed simple properties of the threads and of their labellings. We captured them in some basic models, either totally random or focusing one particular property. We then compared the artificial interaction networks obtained by combining these models to the original ones, and to random ones. It appears clearly that some non-trivial properties of the original network, missed by the usual random models, are captured by the multi-level approach.

Our aim here is not to say that the models we propose are relevant and capture some real-world feature. But we give evidence of the relevance of such an approach to capture, explain and model subtle properties of complex networks, which would be very hard with the classical approach.

We are convinced that this result is very general. Many networks are actually induced by a construction process which can be simply described (and which often relies on the merging of small substructures). Let us cite for instance co-authoring networks, in which authors are linked together if they signed a paper together: each paper induces a clique, which may be seen as responsible for the high clustering [28], [42], and the overall structure of the network is induced by the way these cliques overlap. Modeling such networks by first capturing the redundency between co-authoring relations would certainly make sense. The actor network and co-occurrence networks are also in this case. Going further, many social networks may be seen as the union of ego-centered networks; modeling these small networks and the way they are combined to form the global network is a natural perspective of our work.

Following these remarks, there are at least two clear direction in which our work should be continued. One the one hand, one could certainly use this approach and the models we proposed (or similar ones) to give social interpretations of the observed properties. Indeed, even if we did not discuss this here, the models actually rely on simple social assumptions which we show can be seen as responsible for the properties of the whole network. Analysing this from a social science point of view remains to be done. On the other hand, this approach has the important advantage of relying on very simple models, which makes it possible to *prove* their properties, and their influence on the whole. An analytic study is then possible and would lead to a tightening of theoretical and practical questions.

One may also improve this work by proposing better models for the different levels, or even another multi-level modeling. As already noticed, it is indeed possible to see the data at a wide variety of levels. Some may be relevant depending on the objectives. Likewise, many other statistics could be considered and lead to new insight. As already discussed in Section III, one could also view the network as directed, weighted, and also as evolving during time. There is currently an important lack of methods and tools to tackle the complexity induced by this richer information, but it makes no doubt that it would improve significantly our understanding of the underlying objects and phenomena. As already pointed out, the multi-level formalism has important advantages to tackle this.

Acknowledgments: This work was supported in part by the PERSI (*Programme d'Étude des Réseaux Sociaux de l'Internet*) project and by the GAP (*Graphs, Algorithms and Probabilities*) project.

REFERENCES

[1] Eytan Adar, Li Zhang, Lada A. Adamic, and Rajan M. Lukose. Implicit structure and the dynamics of blogspace. 2004.

[2] Filip Agneessens, Henk Roose, and Hans Waege. Choices of theatre events: p* models for affiliation networks with attributes. *Metodoloski zvezki*, 1(2):419–439, 2004. Short version presented at SunBelt 2002.

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

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

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

[6] Stefano Battiston and Michele Catanzaro. Statistical properties of corporate board and director networks. *European Physics Journal B*, 38:345–352, 2004.

[7] Valérie Beaudouin and Julia Velkovska. Constitution d'un espace de communication sur internet. *Réseaux*, 17(97):121–177, 1999. In French.

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

[9] B. Bollobas. *Random Graphs*. Cambridge University Press, 2001.

[10] Phillip Bonacich. Technique for analyzing overlapping memberships. *Sociological Methodology*, 4:176–185, 1972.

[11] Moses A. Boudourides and Iosif A. Botetzagias. Networks of protest on global issues in Greece 2002-3. 2004. Work in progress – Preprint.

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

[13] Guido Caldarelli, Stefano Battiston, Diego Garlaschelli, and Michele Catanzaro. Emergence of complexity in financial networks. *Lecture Notes in Physics*, 650:399–423, 2004.

[14] Martin J. Conyon and Mark R. Muldoon. The small world network structure of boards of directors. 2004. SSRN preprint, `http://ssrn.com/abstract=546963`.

[15] Wang Dahui, Zhou Li, and Di Zengru. Bipartite producer-consumer networks and the size distribution of firms. 2005. ArXiV preprint `physics/0507163`.

[16] Debian user mailing-lists. `http://lists.debian.org/users.html`.

[17] S.N. Dorogovtsev and J.F.F. Mendes. Evolution of networks. *Advances in Physics*, 51, 2002.

[18] Enron e-mail database. `http://www.cs.cmu.edu/˜enron/`.

[19] P. Erdös and A. Rényi. On random graphs I. *Publications Mathematics Debrecen*, 6:290–297, 1959.

[20] Guler Ergun. Human sexual contact network as a bipartite graph. 2002. ArXiV preprint `cond-mat/0111323`.

[21] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *Proceedings og the internatonal ACM conference SIGCOMM*, pages 251–262, 1999.

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

[23] F. Le Fessant, S. Handurukande, A.-M. Kermarrec, and L. Massoulié. Clustering in peer-to-peer file sharing workloads. In *3-rd International workshop on Peer-To-Peer Systems (IPTPS)*, 2004.

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

[25] Jean-Loup Guillaume, Stevens Le Blond, and Matthieu Latapy. 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)*, 2004.

[26] Jean-Loup Guillaume, Stevens Le Blond, and Matthieu Latapy. 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)*, 2005.

[27] Jean-Loup Guillaume and Matthieu Latapy. 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)*, 2004.

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

[29] Jean-Loup Guillaume and Matthieu Latapy. Complex network metrology. *Complex Systems*, 2005. To appear.

[30] Jean-Loup Guillaume and Matthieu Latapy. Relevance of massively distributed explorations of the internet topology: Simulation results. In *Proceedings of the 24-th IEEE international conference INFOCOM*, 2005.

[31] S. Handurukande, A.-M. Kermarrec, F. Le Fessant, and L. Massoulié. Exploiting semantic clustering in the edonkey p2p network. In *11-th ACM SIGOPS European Workshop (SIGOPS)*, 2004.

[32] Adriana Iamnitchi, Matei Ripeanu, and Ian Foster. Small-world file-sharing communities. *Proceedings of the 23-rd IEEE international conference INFOCOM*, 2004. ArXiV preprint `cs.DC/0307036`.

[33] Jon Kleinberg. Navigation in a small world. *Nature*, 406:845, 2000.

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

[35] PieSpy IRC logger. `http://www.jibble.org/piespy/`.

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

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

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

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

[40] Mark E. J. Newman. Models of the small world. *Journal of Statistial Physics*, 101:819–841, 2000.

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

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

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

[44] Roberto N. Onody and Paulo A. de Castro. Complex network study of brazilian soccer players. 2004. ArXiV preprint `cond-mat/0409609`.

[45] Saverio Perugini, Marcos Andre Goncalves, and Edward A. Fox. A connection-centric survey of recommender systems research. 2003. ArXiV preprint `cs.IR/0205059`.

[46] Garry Robins and Malcolm Alexander. Small worlds among interlocking directors: Network structure and distance in bipartite graphs. *Computational & Mathematical Organization Theory*, 10(1):69–94, 2004.

[47] Camille Roth and Paul Bourgine. Epistemic communities: description and hierarchic categorization. 2005. ArXiV preprint `nlin/0409013`.

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

[49] Brian Uzzi and Jarrett Spiro. Collaboration and creativity: The small world problem. *American Journal of Sociology*, 2005. To appear.

[50] S. Voulgaris, A.-M. Kermarrec, L. Massoulie, and M. van Steen. Exploiting semantic proximity in peer-to-peer content searching. In *10-th IEEE international workshop on Future Trends in Distributed Computing Systems (FTDCS)*, 2004.

[51] Patrick A. Wagstrom, James D. Herbsleb, and Karhleen Carley. A social network approach to free / open source software simulation. *Proceedings of the 1-st international conference on open source systems*, pages 16–23, 2005.

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

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

[54] Kim Young-Choon. A structural analysis on firm-market affiliation networks in the korean system integration industry. *Development and Society*, 27(2), 1998.