Complex Network Metrology

Jean-Loup Guillaume*
Matthieu Latapy†
LIAFA – CNRS – Université Paris 7,
2 place Jussieu, 75005 Paris, France

In order to study complex networks like the Internet, the World Wide Web, social networks, or biological networks, one first has to explore them. This gives a partial and biased view of the real object, which is generally assumed to be representative of the whole. However, until now nobody knows how and how much the measure influences the results.

Using the example of the Internet and a rough model of its exploration process, we show that the way a given complex network is explored may strongly influence the observed properties. This leads us to argue for the necessity of developing a science of metrology of complex networks. Its aim would be to study how the partial and biased view of a network relates to the properties of the whole network.

1. Introduction

Some complex networks of high interest can only be known after an exploration process. This is in particular true for the Internet (interconnection of computers), the World Wide Web (links between pages), social networks (acquaintance relations), and biological networks (brain topology or protein interactions). There have been many studies published on these objects, see for instance [1–11]. Most of them rely on partial views obtained using various, and often intricate, exploration methods. Until now, the approach generally used is to obtain views as large as possible and then assume that they are representative of the whole, see for instance [12–15]. However, except in a few limited cases [12, 16, 17], nobody understood the bias introduced by the partial exploration methods and the influence it may have on the results.

We show here that this bias may be very important, even under some very optimistic assumptions. Using the representative example of the Internet topology, we show how some natural models of the exploration process yield very different views of a given network, which proves that the way one explores a complex network has a strong influence on the properties of the obtained view. We therefore insist on the necessity of

*Electronic mail address: guillaume@liafa.jussieu.fr.
†Electronic mail address: latapy@liafa.jussieu.fr.

developing a theory of complex network metrology. Its aim would be to study how the partial and biased view of a network relates to the properties of the whole network.

Our global approach is the following: we consider a (known) network G, simulate an exploration of this network to obtain a *view* G' of it, and then compare the two objects. The final aim is to deduce properties of G from properties of G'. In this paper, we only make a first step in the direction of this ambitious objective, but we show that it is enough to prove its validity and relevance, which is our aim. In order to do this, we first present the way the Internet topology is explored, then introduce very simple and natural models to simulate this, and finally discuss the obtained results. Let us insist on the fact that this global approach is absolutely general, and may be applied to other cases (like the World Wide Web, social networks, or biological networks) with benefit.

2. Exploring the Internet

Many operators and administrators act on the Internet topology in a totally distributed way. There is no central decision on what is done on the Internet, and no central knowledge of its topology. And yet, it plays an important role in many contexts like the robustness of the network [18].

There are various ways to retrieve some data on the Internet topology from publicly available sources. They give a partial view of the global topology. Moreover, the available information is influenced by many parameters (e.g., economical, technical, or political) which may introduce a bias in the sample we get. This is however the unique method available to know this topology and what we call *exploring the Internet*.

There exist various methods and many heuristics to extensively explore the Internet. We do not enter into the details of these techniques, but will concentrate on one of the main methods. This restriction is motivated both by the fact that very large explorations of the Internet have indeed been conducted using this method (e.g., [4, 12–15]), and that it is quite easy to model, whereas other methods are much less precisely defined.

We concentrate on the exploration of the Internet using the tool `traceroute`. It is a simple program which, used from a *source* computer, gives the path followed by messages from this source to a *destination* computer on the Internet. This path is a set of nodes and links of the network, which can be seen as a small part of the Internet topology. Using this tool extensively, one can obtain large parts of the whole topology.

Note that, in order to use `traceroute`, one has to run the program on the source computer. On the contrary, nothing specific is needed at the destination and so one can choose any destination. Therefore, if one uses traceroute to explore the Internet, the number of sources used is generally very limited (typically a few dozens) whereas the number of destinations may be huge (typically several hundreds of thousands) (e.g., [12, 13, 15]). Note also that, if one explores the Internet from one source, one cannot obtain a perfect view of the whole, even if traceroute is used to every possible destination. Indeed, there are some links which will never be crossed by any message from the source. Moreover, due to bandwidth, knowledge, and time limitations, one can never use traceroute to every possible destination. How many destinations should one consider? How many sources are needed? Until now, no one has any idea of the answers to these questions, but we propose a step towards them.

3. Modeling

We want to simulate an exploration process. In order to do this, we first need a network to explore. There are several natural choices for this. One can, for instance, obtain the real topology of a large computer network provided by a firm. One can also use one of the various models proposed to generate random networks (e.g., [19–26]). It has been shown recently that the Internet topology, like many other complex networks, has specific statistical properties [4]. However, in this paper we are mostly concerned with the exploration process. Therefore, we choose the most simple and well-known model of random networks [27, 28] to generate the topology that is explored: the Erdös and Rényi random graph model. This model has two parameters: the number of nodes n, and the probability of the existence of any link p. A network is then generated by considering that each possible pair of nodes is linked with probability p. This gives an expected number of links $m = p \cdot n \cdot (n-1)/2$. Note that this model is not very realistic, but it is sufficient for the purpose of this paper.

The traceroute tool gives the path followed by messages from a source to a destination. Until now, very little is known about the properties of such paths, see [29] and the references therein. For instance, one may suppose that the aim of network protocols is to deliver information efficiently, and so that the paths followed are the shortest paths (paths of minimal length). It is however known that this is not always the case, but no precise information is currently available on how they differ from shortest paths [29]. Moreover, there exist in general many shortest paths for a given pair of computers, and there is no *a priori* reason for traceroute to give one of them rather than another. Finally, the paths change during time but again very little is known on their dynamics.

With the current state of our knowledge, designing a realistic model of traceroute is therefore impossible. The assumption usually made is that traceroute always gives a shortest path, which is actually sufficient for our current aim. We also consider that, during the exploration process, one may use traceroute many times, which leads to the discovery of all the shortest paths between given sources and destinations.

We have a model to generate the network to explore, and some models for the traceroute tool. We now need a model for the exploration process itself. As already noted, we suppose that it only relies on traceroute. But this is not sufficient: we must say how we will choose sources and destinations, and how many of them will be considered. With our aim being to show that the exploration method may influence the obtained view of the actual network, we consider several realistic models of the exploration. Again, we only consider the simplest ones, which are sufficient for our purpose. Since it is the case in practice, we suppose that the exploration process is based on one or a few sources, and uses many or all of the possible destinations. Moreover, we suppose that the sources and destinations are chosen randomly, which makes sense since the networks explored are totally random and so all the nodes play similar roles.

Let us insist on the fact that, to make a complete study of the influence of the exploration process on the view we obtain, one would actually have to consider many models, both for the network to explore, for the traceroute behavior, and for the exploration method. Therefore, one obtains several dozens of triples of models to consider, and for which experiments and comparisons should be conducted. However, this is not our aim here. We only want to show that the exploration method indeed influences the results. To achieve this, as shown in the following, it is sufficient to consider a few simple cases.

Finally, the models we use are very simple. The network to explore is produced by the classical random network model, which gives a network of n nodes where each link exists with probability p. We always suppose that traceroute gives shortest paths, but consider both the case where it gives one shortest path and the case where it gives all of them. Finally, we consider a varying number of sources and destinations from one to a few sources and many to all destinations, which reflects the values used in practice. We explained why all these choices are reasonable considering our aim, but clearly many others would be relevant too.

All of the values plotted were averaged over 1000 instances. The variance is in general negligible, as shown later in Figure 2. The shortest path computations are done using breadthfirst search.

4. How much do we see?

We now consider a random network G in which each link exists with probability p. We make explorations of G using various models. We first consider only one source, chosen at random, and then consider the

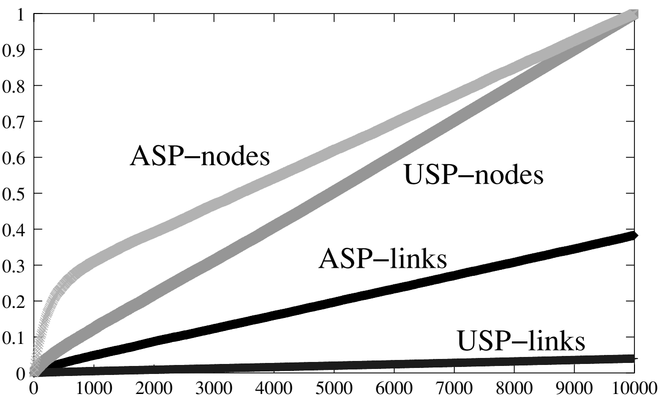

Figure 1. Ratio of the total number of nodes and links discovered during an exploration, as a function of the number of destinations. These plots correspond to a random network with $n = 10,000$ and $p = 0.005$, which gives an average degree in accordance with what is generally assumed for the Internet topology.

case with several sources. All the experiments are conducted with two models of `traceroute`, the USP model (where we discover a unique shortest path between each pair of source and destination), and the ASP model (where we discover all the shortest paths for each pair). The plots are averaged over 1000 runs.

4.1 Unique source

Let us denote by $G_u(x)$ the view of G obtained from a given source if we consider x random destinations, with the USP model for `traceroute`. Let $n_u(x)$ be the number of nodes of this view, and $m_u(x)$ its number of links. Similarly, we introduce $G_a(x)$, $n_a(x)$, and $m_a(x)$, which are the results obtained with the ASP model for `traceroute`. The plots of these functions given in Figure 1 show how much of the network we obtain, both in terms of nodes and links, as a function of the number of destinations.

At various points, these plots fit well with our intuition. First, when we consider very few destinations, we obtain a very small part of the network. Then, if the number of destinations grows, we see more and more. Finally, we can see all of the nodes when we consider each of them as a destination.

There are however a few remarkable facts. Both $n_u(x)$ and $n_a(x)$ grow rapidly and reach a critical point where they start a linear growth, but the initial growth of $n_a(x)$ is much more rapid than that of $n_u(x)$. On the contrary, $m_u(x)$ and $m_a(x)$ grow linearly from the beginning, but the maximal values they reach, $m_u(n)$ and $m_a(n)$, remain surprisingly low. This means that the exploration misses many links, even if we consider all possible destinations, which indicates that the obtained view is very incomplete. This is even more surprising when we consider the optimistic case where all of the shortest paths are discovered, and all of the nodes are used as destinations.

These behaviors are similar for any values of n and p (the plots presented in Figure 1 always have the same shape). However, the maximal value reached by $m_u(x)$ and $m_a(x)$, that is, the maximal proportion of discovered links, varies with the probability p for the existence of any link. To know how p influences these values, let us study the proportion of links discovered using one source and all possible destinations, as a function of p. They are plotted in Figure 2 for the two models of `traceroute` considered.

The two plots have some properties in common which can be easily explained. First notice that below a certain value of p, the network is not connected (it is composed of many independent parts) [28]. Therefore, below this threshold, any exploration using a small number of sources will give a very small part of the whole. When the network becomes connected, it is almost a tree, in which there is a unique path from the source to each node. Therefore, the two exploration methods we consider discover almost all of the links, which corresponds to the maximal values reached by the plots in Figure 2. On the other hand, when p is almost 1, then almost every possible link exists, and so almost every node is at distance 1 from the source. Therefore, the obtained view, with both the USP and ASP models, is almost a star. It therefore contains almost $n-1$ links, which when compared to the total number of links, almost $n \cdot (n-1)/2$, is negligible.

The plot for the USP model is easy to understand. Indeed, the exploration using this model gives a tree (it has no cycle), and therefore it contains exactly $n-1$ links if p is above $\log(n)/n$ since in this case the network is almost surely connected. The expected total number of links being itself $m = p \cdot n \cdot (n-1)/2$, the ratio between the number of links discovered during the exploration and the total number of links is then $(n-1)/m = 2/p \cdot n$. When p grows, this ratio decays as $1/p$, which is confirmed by the simulation.

On the contrary, the irregular shape of the plot for the ASP model (which we name the *camel* plot) is very surprising: it has many peaks and valleys of high amplitude, which have no obvious interpretation. There is however a natural explanation of this shape, which comes from specific properties of the exploration.

4.2 The camel plot

Let us first characterize the links missed during the exploration. If a link is on a shortest path from the source to any other node then it is discovered, since all shortest paths to all nodes are discovered.

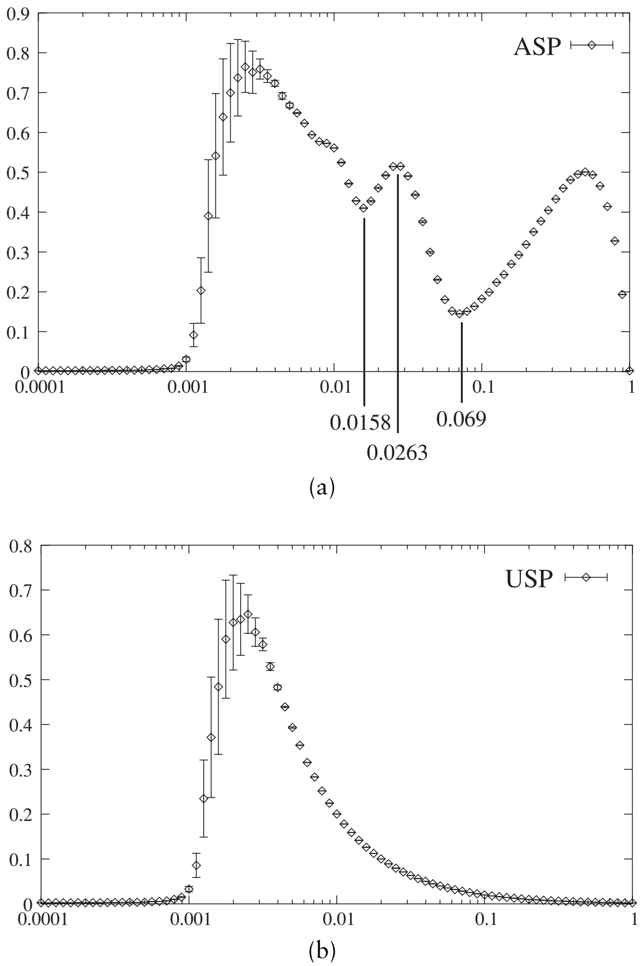

Figure 2. Proportion of discovered links (one source, all destinations) as a function of p for random graphs with $n = 1000$. (a): ASP; (b): USP. The plots are the average over 1000 instances, and the variance is displayed (it is negligible everywhere except at the connectivity threshold). The plot obtained in the ASP case has a surprising shape, leading to the name *camel* plot.

Conversely, if a link is discovered during the exploration, it has to be on a shortest path. Therefore, we miss the links that are not on a shortest path from the source to any other node. These links are exactly the ones between nodes at equal distance from the source. In other words, the function plotted in Figure 2 is nothing but m minus the number of links between nodes equidistant from the source, over m.

Now let us consider the number of such links. To do this, we consider the distribution of the distances from the source. As shown in Figure 3,

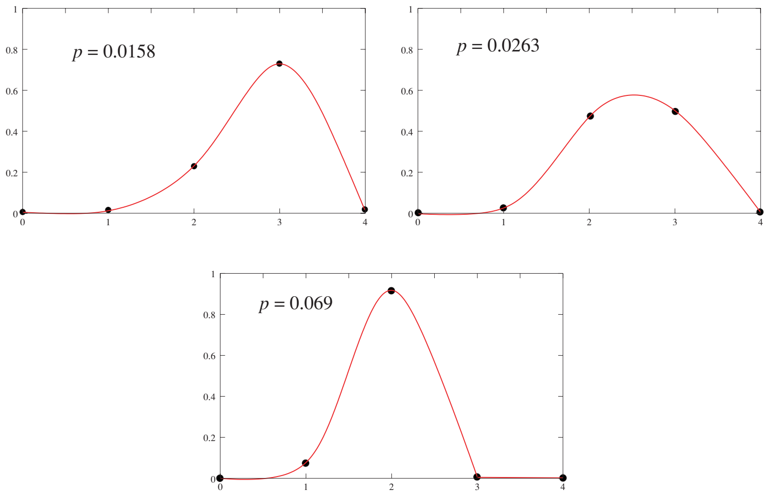

Figure 3. Distance distribution from the source for random networks (n = 1000 nodes) with various links densities p. The distribution is centered around the mean distance, which decays smoothly as p grows.

this distribution is centered around its mean value, which decays when p grows. This is not surprising, and notice that it has the same global shape independently of p. So, how can it help in understanding the camel plot? The point is that we have to consider the discrete distribution of the distances from the source, also displayed in Figure 3. Since distances are integers, these discrete distributions are the actual distributions. But when we consider a discrete distance distribution, two cases may occur: the mean distance (or the distance for which the continuous distribution is maximal) can be close to an integer or it can be well-centered between two integers. In the first case, almost all of the nodes will be at this distance from the source, while in the second case almost half of them will be at some distance from the source and the other half at this distance plus one. These two cases are illustrated in Figure 3 (near-integer with $p = 0.0158$ and $p = 0.069$, and well-centered for $p = 0.0263$). Recall that we miss the links between nodes at the same distance from the source. Therefore, when most nodes are at the same distance from the source, we miss many links, much more than in the other case. Since the average distance decays when p grows, there is an alternate series of such phases, which correspond to the peaks and valleys of the camel plot.[1]

[1]We checked this by computing the distance distributions of graphs and then the number of links between two nodes at the same distance from the source. The obtained results exactly fit the camel plot.

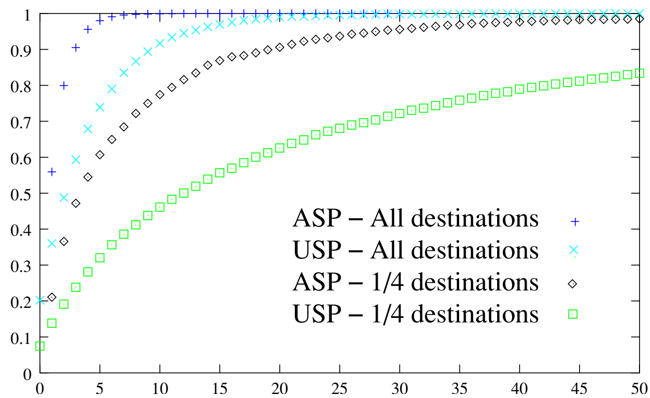

Figure 4. Variation of the amount of discovered links as a function of the number of sources, in two cases: if all the nodes are destinations, and if only a quarter of them are. This plot corresponds to $n = 2000$ and $p = 0.005$, which leads to the conclusion that $50 = 2.5\%$ of the nodes should be used as sources. This is much more than usually done for the Internet.

These first results clearly show that even very simple properties like the ratio of discovered links cannot be easily derived from a partial view of the network. Indeed, the efficiency of the exploration method varies a lot with network properties like density of links, and, more surprisingly, small variations in these properties may have a strong impact on the exploration significance.

4.3 Several sources

Until now, we have restricted ourselves to explorations using only one source. However, in practical cases, one generally uses a few sources. We investigate here how this may influence the quality of the view obtained. Again, we only concentrate on the ratio of the total number of discovered links, which previous remarks have shown to be essential.

Figure 4 shows the evolution of this ratio when the number of sources increases. We first consider the two uppermost plots, which correspond to the cases where we use all of the possible destinations. As expected, the quality of the view grows rapidly with the number of sources, and one may even be surprised by the rapidity of this growth. Despite the roughness of our model for Internet exploration, one may consider this plot as good news since it indicates that many sources are not needed to obtain accurate views of the network. This is important since it is very difficult (and never done) to use many sources in practice.

However, the assumption that all the nodes of the network serve as destinations is very rough. It is difficult to give an estimation of the number of nodes which actually contribute as destinations, but we can, for instance, suppose that only a quarter of them do, which is already huge. We then obtain the two lower plots in Figure 4. Whereas the upper plots made us relatively optimistic, these show that a large number of sources are necessary to obtain an accurate view of the whole.

All these experiments cannot lead to conclusions concerning the exploration of the Internet itself. They do show however, that a very reasonable hypothesis (in the limited state of our current knowledge) for the exploration process leads to qualitatively different results. This is important evidence for taking the exploration process into account.

5. Conclusion

In this paper, we considered the simplest possible question concerning the quality of a network view obtained by exploring a real network: the total number of nodes and links obtained. Making natural variations on the way we model the Internet exploration, we show that this amount varies a lot and is difficult to estimate.

Other properties, like the degree distribution or the clustering, are also biased by the exploration process. Moreover, as discussed, many models are possible for the exploration process, and we presented only a few simple ones here. However, the results we have presented are representative of what happens in all other cases and are sufficient for our purpose. This, added to their simplicity, is why we chose them to illustrate our arguments.

Let us insist once more on the fact that the results presented here do not provide any information on the Internet topology itself. They do not even give any information on how much, and how, the known results of the Internet topology are biased by the partial exploration process. Instead, they give evidence for the fact that this bias exists and may be very important. This fact is very general and can be shown in a similar fashion for the World Wide Web graph, various social or biological networks, and other complex networks.

We therefore argue that there is a need for the development of a new area of scientific activity, focused on complex network metrology. Results in this area are highly needed and would make it possible to give rigorous results on a variety of complex networks that cannot be studied directly. We suspect that this is actually the case for most complex networks.

References

[1] R. Albert, H. Jeong, and A.-L. Barabási, "Diameter of the World Wide Web," *Nature*, **401** (1999) 130–131.

[2] 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," *Ninth International World Wide Web Conference / Computer Networks*, **33**(1–6) (2000) 309–320.

[3] K. Efe, V. Raghavan, C. H. Chu, A. L. Broadwater, L. Bolelli, and S. Ertekin, "The Shape of the Web and Its Implications for Searching the Web," in *Proc. Int. Conf. Advances in Infrastructure for Electronic Business, Science, and Education on the Internet*, 2000 (Scuola Superiore Guglielmo Reiss Romoli, 2000).

[4] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On Power-law Relationships of the Internet Topology," in *ACM SIGCOMM*, pages 251–262, 1999.

[5] H. Jeong, B. Tombor, R. Albert, Z. Oltvai, and A.-L. Barabási, "The Large-scale Organization of Metabolic Networks," *Nature*, **407** (2000) 651.

[6] J. M. Kleinberg and S. Lawrence, "The Structure of the Web," *Science*, **294** (November 2001) 1849–1850.

[7] F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley, and Y. Aberg, "The Web of Human Sexual Contacts," *Nature*, **411** (2001) 907–908.

[8] A. Medina, I. Matta, and J. Byers, "On the Origin of Power Laws in Internet Topologies," in *ACM Computer Communication Review*, **30**(2) (April 2000).

[9] S. Milgram, "The Small World Problem," *Psychology Today*, **1** (1967) 61–67.

[10] S. Milgram, "The Small World Problem," in *The Individual in a Social World: Essays and Experiments*, edited by J. Sabini and M. Silver (McGraw Hill, New York, second edition, 1992).

[11] Small World Project, http://smallworld.columbia.edu/project.html.

[12] R. Govindan and H. Tangmunarunkit, "Heuristics for Internet Map Discovery," in *IEEE Conference on Computer Communications 2000*, pages 1371–1380, Tel Aviv, Israel, March 2000.

[13] Y. Hyun, A. Broido, and K. Claffy, "Traceroute and BGP AS Path Incongruities," http://www.caida.org/outreach/papers/2003/ASP/.

[14] D. Magoni and J.-J. Pansiot, "Analysis of the Autonomous System Network Topology," *ACM SIGCOMM Computer Communication Review*, **31**(3) (July 2001) 26–37.

[15] N. Spring, R. Mahajan, and D. Wetherall, "Measuring ISP Topologies with Rocketfuel," in *Proceedings of ACM SIGCOMM '02*, August 2002.

[16] A. Lakhina, J. Byers, M. Crovella, and P. Xie, "Sampling Biases in IP Topology Measurements," in *IEEE Conference on Computer Communications*, 2003.

[17] P. De Los Rios, "Exploration Bias of Complex Networks," in *Proceedings of the Seventh Conference on Statistical and Computational Physics Granada* (American Institute of Physics Conference Proceedings, Volume 661, 2002).

[18] R. Albert, H. Jeong, and A.-L. Barabási, "Error and Attack Tolerance in Complex Networks," *Nature*, **406** (2000) 378–382.

[19] R. Albert and A.-L. Barabási, "Emergence of Scaling in Random Networks," *Science*, **286** (1999) 509–512.

[20] R. Albert and A.-L. Barabási, "Statistical Mechanics of Complex Networks," *Reviews of Modern Physics*, **74** (2002) 47.

[21] S. N. Dorogovtsev and J. F. F. Mendes, "Evolution of Networks," *Advances in Physics*, **51** (2002) 1079–1187.

[22] J.-L. Guillaume and M. Latapy, "Bipartite Structure of *All* Complex Networks," *Information Processing Letters (IPL)*, **90**(5) (2004) 215–221.

[23] J.-L. Guillaume and M. Latapy, "Bipartite Graphs as Models of Complex Networks," in *Lecture Notes in Computer Sciences (LNCS), Proceedings of the First International Workshop on Combinatorial and Algorithmic Aspects of Networking (CAAN)*, 2004.

[24] M. E. J. Newman, "The Structure and Function of Complex Networks," *Society for Industrial and Applied Mathematics Review*, **45**(2) (2003) 167–256.

[25] M. E. J. Newman, D. J. Watts, and S. H. Strogatz, "Random Graph Models of Social Networks," *Proceedings of the National Academy of Sciences, USA*, **99**(Suppl. 1) (2002) 2566–2572.

[26] D. J. Watts and S. H. Strogatz, "Collective Dynamics of Small-world Networks," *Nature*, **393** (1998) 440–442.

[27] P. Erdös and A. Rényi, "On Random Graphs: I, *Publicationes Mathematicae Debrecen*, **6** (1959) 290–297.

[28] B. Bollobás, *Random Graphs* (Academic Press, London, 1985).

[29] T. Friedman, M. Latapy, J. Leguay, and K. Salamatian, "Describing and Simulating Routes on the Internet," in *Proceedings of the Fourth International Federation for Information Processing International Conference on Networking*, 2005.