Relevance of Massively Distributed Explorations of the Internet Topology: Qualitative Results

Jean-Loup Guillaume, Matthieu Latapy, Damien Magoni

Relevance of Massively Distributed Explorations of the Internet Topology: Qualitative Results [1]

Jean-Loup Guillaume [2], Matthieu Latapy [2]
and Damien Magoni [3]

***Abstract*— Internet maps are generally constructed using the `traceroute` tool from a few sources to many destinations. It appeared recently that this exploration process gives a partial and biased view of the real topology, which leads to the idea of increasing the number of sources to improve the quality of the maps. In this paper, we present a set of experiments we have conducted to evaluate the relevance of this approach. It appears that the statistical properties of the underlying network have a strong influence on the quality of the obtained maps, which can be improved using massively distributed explorations. Conversely, some statistical properties are very robust, and so the known values for the Internet may be considered as reliable. We validate our analysis using real-world data and experiments, and we discuss its implications.**

***Index Terms*— Internet topology, graphs, metrology, active measurements.**

INTRODUCTION.

Due to its fully distributed construction and administration, mapping the Internet (in terms of IP routers and IP-level links between them) is a challenging task. It is however essential to obtain some information on its global shape. Indeed, it plays a central role in key problems like network robustness, see for instance [43], [4], [14], [15], simulation of future protocols and uses, see for instance [41], and many others.

Exploring the Internet topology is a research problem in itself, see for instance [26], [28], [39], [57], [62]. Indeed, many difficulties (like the identification of the multiple interfaces of a same router) arise when one wants to map the Internet. Various techniques and methods have been introduced to achieve this goal. Some of them are very subtle, but current explorations still rely on the extensive use of the `traceroute` tool: one collects routes from a given set of sources to a given set of destinations, and then merges the obtained paths. Some post-processing is generally necessary to clean the obtained data, but we do not enter in these details here.

Two points are particularly important in the scheme sketched above. First, it must be clear that the image we obtain from the network is *partial* (except if the number of sources and destinations is huge, we certainly miss some nodes and some links) and may be *biased* by the exploration process (some properties of the obtained map may be induced by the way we explore the network, not by the network itself). Second, the number of sources cannot be increased easily, whereas one can take as many destination as one wants. Indeed, one needs direct access to the sources in order to run the `traceroute` tool, whereas one only needs the IP addresses of the destinations. In the case of [26] for instance, which is one of the largest explorations currently available, only a few dozens of sources are used whereas there are several hundreds of thousands destinations.

Recently, several researchers conducted experimental and formal studies to evaluate the accuracy of the obtained maps of the Internet [1], [13], [17], [18], [31], [32], [33], [35], [52], [56]. All these studies use simple models of networks and `traceroute` but they all give good arguments of the fact that the currently available maps of the Internet are very incomplete, and that there probably is an important bias induced by the exploration process.

In order to improve these maps, several researchers and groups now propose to deploy massively distributed measurement tools [25], [53], [55]. The basic idea is that dramatically increasing the number of sources would significantly improve the quality of the obtained maps. Our central aim in this paper is to rigorously evaluate the relevance of this approach.

To achieve this, we conduct an extensive set of experiments designed as follows, according to the natural methodology already used for instance in [13], [31], [35]. We consider a graph G representing the network to explore. We then simulate the exploration process and obtain this way a (partial and biased) view G' of the original graph. We then compare G' and G to evaluate the quality of this view. We process this simulation using all the possible numbers of

[1] A reduced conference version of this contribution [32] has been accepted for publication in the proceedings of the 24-th IEEE international conference INFOCOM, 2005.

[2] LIAFA – CNRS – Université Paris 7 – 2 place Jussieu, 75005 Paris, France. (guillaume,latapy)@liafa.jussieu.fr

[3] LSIIT – CNRS – Université Strasbourg 1 – UFR de Math-Info – 7, rue René Descartes F-67084 Strasbourg, France

sources and destinations, which makes it possible to study the impact of these numbers on the accuracy of the obtained view. Likewise, we take a variety of graphs as models of the network, with very different properties, in order to investigate their influence on the exploration process and how much this process is able to capture them. We also study an important real-world data set which makes it possible to evaluate the relevance of the simulations.

The paper is organized as follows. First we define the statistical properties of networks relevant to our study, we present the models we use and discuss our methodology (Section I). Then we present and analyze the results of our simulations on various models and statistical properties (Sections II–IV). We show how our approach can be used to design efficient exploration strategies by choosing appropriate sources and destinations in Section VI. Section VII is devoted to the comparison of our results with real-world data and experiments, which makes it possible to identify the most meaningful simulations and to evaluate our hypotheses. Finally we present our conclusions and discuss them.

I. Preliminaries.

A network topology can naturally be represented by a graph. For our purpose, the graph does not need to be weighted nor directed. A route in the network, as given by the `traceroute` tool, is a path in the corresponding graph. For a few years, a strong effort has been made to discover the topology of the Internet at IP and/or router level by extensive use of `traceroute` and other tools (BGP tables, source routing, etc). See for instance [12], [24], [26], [50].

The obtained maps give much information on the global shape of the Internet. In particular, they gave evidence of the fact that the Internet topology has some statistical properties which makes it very different from the models used until then, see for instance [10], [24]. This induced an intense activity in the acquisition of such maps [26], [28], [50], in their analysis [24], [59] and in the accurate modeling of the Internet [9], [40], [63], [64]. See [51] for an in-depth survey.

Our analysis of the exploration process will be based on these statistical properties and these models, which we present below. We also need to model the `traceroute` tool and the exploration process, which we also discuss in this section. Finally, we present our methodology, and explain how our results should be read.

Statistical properties

The Internet, at router level, is composed of several millions of nodes and dozens of millions of links. Let N denote its number of nodes and M its number of links.

It is well known, and quite intuitive, that the density of the Internet graph is low: the number of existing links over the number of possible ones, $\frac{2 \cdot M}{N \cdot (N-1)}$, is low. In other words, the average degree k of the nodes (their average number of links), *i.e.* $k = \frac{2 \cdot M}{N}$, may be viewed as a constant independent of the size of the network.

A less known point is that the average distance (length of a shortest path between two nodes) is low. It typically scales as $\log(N)$. This is however not surprising, since it is an essential objective of the design of the network, and since it is actually very natural for any graph with some amount of randomness to have a low average distance, see for instance [8], [37], [48]. In some specific cases, the average distance can even scale as $\log \log(N)$ or be bounded by a constant independent of n [11], [60], [61], [49].

On the contrary, although it is now well understood, the fact that the degree distribution of the Internet graph is very heterogeneous has been a surprise [24]. Indeed, the proportion p_k of nodes of degree k might be approximated by a power of k: $p_k \sim k^{-\alpha}$ with $\alpha \simeq 2.5$. Intuitively, this means that most nodes have a low degree but there exists some nodes with (very) high degree. Such graphs are said to be *scale-free*.

Another important statistical property measured on the Internet is its clustering C defined as $C = \frac{N_\Delta}{N_\vee}$, where N_Δ is the number of triangles (three nodes with three links) in the network and $N_\vee$ is the number of connected triples (three nodes with at least two links)[4]. In other words, C is the probability that two nodes are connected together, given that they are both connected to a same third, which gives a measure of the local density of the graph. The clustering of the Internet is high, considered as a constant independent of N.

[4]There are several definitions for the notion of clustering coefficient, which all have their own advantages and drawbacks. They are all aimed at capturing the local density of graphs, and would serve our purpose equivalently.

Modeling networks

The basic model for networks is the Erdos and Rényi (ER) random graph model [8], [22]. In an ER graph with n nodes, each of the $\frac{n\cdot(n-1)}{2}$ possible links exists with a given probability p. Equivalently, an ER graph is constructed from n nodes by choosing $m = p \cdot \frac{n\cdot(n-1)}{2}$ links at random. Notice that an ER graph contains a giant component as soon as the average degree is greater than 1 [8]. In the following this condition is always fulfilled and generally the graph itself is fully connected.

In such a graph, the average distance grows as $\log(n)$ [8] as long as p is high enough. However, the clustering is small (it tends to zero when n grows), and the degree distribution follows a Poisson law ($p_k \sim e^{-\alpha}\frac{\alpha^k}{k!}$). This implies in particular that all the nodes have a degree close to the average. Therefore, although this model can be considered as relevant concerning the average distance, it misses the two other main properties of the Internet.

An important step was made when Albert and Barabási (AB) introduced their model based on *preferential attachment* [2], [20]. In this model, nodes arrive one by one and choose k neighbors among the existing ones with a probability proportional to their degree. The degree distribution of the nodes in the obtained graphs follow a power-law with an exponent -3 (it is possible to modify this exponent in others models using preferential attachment). The average distance of such a graph is logarithmic in the number of nodes, but the clustering is low.

This model has been modified to give highly clusterized graphs: in the Dorogovtsev and Mendes (DM) model [19], nodes arrive one by one but at each step one chooses a random *link* $\{u, v\}$ and the new node is linked to both u and v. This implies that a node is chosen with a probability proportional to its degree. Therefore, the preferential attachment principle is hidden in this model, which induces the fact that DM graphs have a power-law degree distribution. Moreover, since one forms a triangle at each step, they have a high clustering.

It is also possible to sample a random graph with a prescribed degree distribution using the Molloy and Reed [5] (MR) model [38], [46], [47]. This gives graphs with exactly the wanted degree distribution, but with low clustering. [7], [38], [46], [47].

Finally, the Guillaume and Latapy (GL) model [29], [30], based on bipartite graphs, gives graphs with power law degree distributions and high clustering, by sampling graphs with prescribed distribution of clique (complete sub-graph) sizes.

Model	Density	Distance	Degree	Clustering
ER	YES	YES	NO	NO
AB	YES	YES	YES	NO
MR	YES	YES	YES	NO
DM	YES	YES	YES	YES
GL	YES	YES	YES	YES

TABLE I

CHARACTERISTICS OF THE MODELS WE USE IN THIS PAPER CONCERNING THE MAIN STATISTICAL PROPERTIES.

The performance of these models are summarized in Table I. They are currently the most widely used for the realistic modeling of clusterized scale-free networks and have all their own advantages and drawbacks. In particular, the parameters are different from one model to another: the main parameter for ER and AB models is the average degree, and the others properties of these models (the degree distribution for instance) are consequences of the construction process itself. Likewise, the original DM model has no parameter but the size of the generated graph and once again, the properties of this model are contained in the construction process. Finally, MR and GL models are defined using the degree distributions one wants to obtain, and most of the properties (including the average degree) are consequences of these distributions. Therefore, depending on the targeted property (degree distribution, clustering, etc), one will use one model rather than another.

These models have been considered as building blocks for more complex models. See [3], [9], [19], [23], [34], [45], [51], [58], [63] for a description of some of these.

In the results we present here, our aim is to give evidence of the impact of the network properties on the efficiency of a shortest-paths based exploration. In most cases, the results do not vary qualitatively between the AB and the MR models on the one hand (which have a power-law degree distribution and no clustering), and between the DM and the GL ones on the other hand (both power-law degree distribution and clustering). We will therefore mainly present

[5] Despite it has been introduced in [6] and studied by Bollobas in [7], this model is commonly refferred to as the *Molloy and Reed* model since these authors made it popular in their contributions [46], [47]. We will follow this convention here.

results on ER, AB and DM models, except in Section VII where it is particularly relevant to use MR and GL ones.

Modeling `traceroute` *and the exploration*

In this paper, we will make the classic assumption [13], [35] that a route as obtained by `traceroute` is nothing but a shortest path between the source and the destination. It is known that this is not always true, see for instance [33], [36]. However, this choice is motivated by the two following points:

- this approximation has little influence, if any, on our results, which we will demonstrate in Section VII,
- and realistic modeling of routes is nowadays a challenging issue for which no better solution usable in our context is known [33], [36].

Moreover, let us emphasize on the fact that we will make an intensive use of route simulations, which makes it crucial to be able to process them very efficiently. To this respect, our assumption has important advantages.

Since there may be many shortest paths between two nodes, this is not sufficient to properly define a model of `traceroute`. At a given moment, the route followed by a packet when a given router routes it to a destination will always be the same independently of the sender. This may have an influence on the quality of the exploration process, therefore we included it in our model of `traceroute`: we always follow the same shortest path (initially chosen randomly) between any two nodes. In [35] a similar model of `traceroute` based on shortest-paths has been introduced.

We now have a precise model of routes as viewed by `traceroute`. But we also need a model for the exploration process. We considered two points of view: in the first one we suppose that we make a *snapshot* of the network, and in the second one we suppose we make a *long-time* exploration. This leads respectively to the *unique shortest path* (USP) model, and to the *all shortest paths* (ASP) one: we either see only one route for any given source and destination, or we see all the possible ones. The ASP model should not be considered as a realistic model, since one cannot expect to get all shortest-paths even within a long period of time (in such a long time, the network is very likely to evolve). However it can be considered as a best case procedure when dealing with shortest-paths or as an upper bound on the amount of information one can expect from a shortest-paths based exploration. The actual quality of such an exploration lies somewhere in between USP and ASP.

We also conducted experiments using other models (random shortest path, several shortest paths but not all, etc), and the results did not qualitatively vary, so we do not detail them here.

Finally, we generally consider a set of sources and a set of destinations, and make the exploration using each possible couple of source and destination in these sets. Such a study has already been conducted on real data in [5], where the authors have defined this exploration scheme as a (k, m)-`traceroute` study (the exact definition appears later in [35]), where k is the number of sources and m the number of destinations chosen at random. Then all `traceroute` are performed between the sources and the destinations. We are going to use a similar approach in the following.

Methodology and grayscale plots

Our global approach is as follows:

1) generate a graph G using a given model with some known properties,
2) compute a view G' of G using a given model of the exploration process and a set of sources and of destinations, and
3) compare the statistical properties of G' to the ones of G.

This methodology is very natural, and has already been used for instance in [13], [31], [35].

Let us insist on the fact that we seek *qualitative* results only: we want to know how qualitative properties of the network influences the properties we observe during an exploration process, and how reliable are the obtained maps with respect to some statistical properties. It makes no sense to interpret quantitatively the results obtained with the kind of approach we use here. On the contrary, by the simplicity of the models and of the properties we use, we obtain evidences of the fact that some properties play a fundamental role in the exploration whereas others may be neglected.

In the method sketched above, the third point (comparison of the original graph with the view we obtain) is a difficult task. It generally leads to a huge amount of plots which one has to compare. To help

in this, we will make an extensive use of grayscale plots which we define as follows (see Figure 8 for some easily readable examples).

For a graph G with N nodes, we consider a square of size $N \times N$. Each point (x, y) of the square corresponds to a view G' of G using x sources and y destinations with a given model of the exploration process (the point $(0, 0)$ is in the lower left corner). The point is drawn using a grayscale representing the value of the non-negative real-valued statistical property p under consideration: from black for $p = 0$ to white for the maximal value of p (which might be greater than the real value).

Therefore, in these plots, the point $(0, 0)$ is always black (we do not see anything using zero sources and zero destinations and in this case all the properties we will consider are null) and the point (N, N) has the grayscale corresponding to the value of p for the original graph G (when every node is a source and a destination, we see everything: $G' = G$). The points darker than the point (N, N) correspond to conditions where the value of p is underestimated, whereas points clearer correspond to conditions where it is over-estimated. The gray variation is linear: if a dot is twice darker than another dot, then the associated value is twice as large.

Notice that each point of such a plot corresponds to a graph G', and computing such plots is computationally expensive. Therefore, is it important to efficiently compute them and to keep N quite low. We conducted experiments with $N = 10^3$, $N = 10^4$ and $N = 10^5$ typically, and, whereas some finite size effects are visible on small graphs ($N = 10^3$), these effects disappear for graphs of size $N = 10^4$ and more. This is why we will present plots for this value of N in general.

Finally, to improve the grayscale plots readability, we added on each such plot the 0.25–, the 0.50–, the 0.75– and the 0.99–level lines, where the l–level line is defined as the set of points where the value of p over its maximal value is between $l - 0.01$ and $l + 0.01$. These lines are often a precious help in the interpretation of the grayscale plots. See Figure 8 and the rest of the paper for examples.

II. Proportion discovered

In this section, we focus on the most basic statistical properties of an exploration, namely the proportion of discovered nodes, the proportion of discovered links, and the quality of the evaluation of the average degree. We present the relevant results on the ER, the AB, the MR and the DM models, and we explain which parameters have a strong influence on these results.

Notice that results using similar approach have been obtained in [5], however our explorations are processed on random graphs instead of real data, the aim being to highlight the parameters of the models and therefore the characteristics of the graphs which influence the efficiency of the exploration.

All possible destinations and few sources

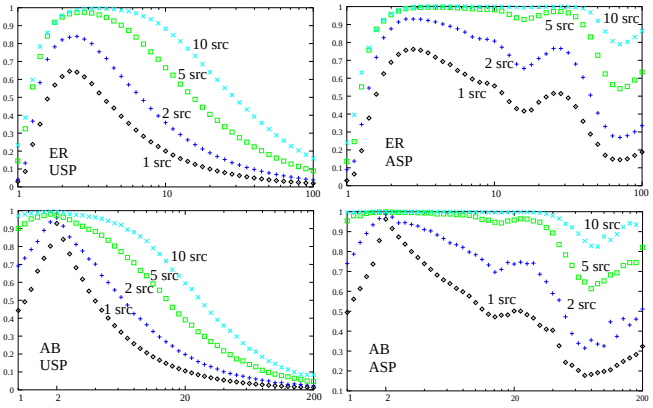

Fig. 1. Proportion of discovered links versus average degree in an ER graph (first row) and in an AB graph (second row) when one uses a USP exploration (left column) or an ASP one (right column). Plots are given for various (small) numbers of sources (namely 1, 2, 5 and 10). All the nodes are taken as destinations. $N = 10^4$.

Let us first study what happens when the number of sources grows but stays small (all the nodes are destinations, therefore we discover all of them). We plot in Figure 1 the proportion of discovered links in several cases, as a function of the (real) average degree for ER and AB graphs (the only ones for which the average degree is a basic parameter). This makes it possible to check some natural intuitions: the quality of the view grows with the number of sources, and it is better for ASP than for USP. Notice however that as the average degree grows, the number of (shortest) paths between two given nodes grows rapidly. Therefore, the ASP exploration becomes more efficient than USP, which fails in discovering many links.

As already explained in [31], the fluctuations in the ASP plots, which may seem surprising, are due to the fact that the missed links are exactly the ones between two nodes at the same distance from the source (such a link cannot be on a shortest path from the

source, and all others are). Therefore, when most nodes are at distance 2 from the source (for instance when the average degree is 69 on a 10^3 nodes ER graph), many links are between them, are therefore are missed (which leads to a hole in the curve). On the opposite, when there are as many nodes at distance 2 from the source as at distance 3 (typically when the average degree is 26), then we miss only few links (and there is a bump on the curve).

Finally, there is no significant difference between the behavior of ER and AB graphs. These plots also give evidence for the fact that, for ER and AB graphs with low average degree, only a few sources are sufficient if the number of destinations is large. The main reason is that with a (very) low average degree, the graph is either non-connected or nearly a tree. In this last case, the graph is obviously easy to discover. The density of the graph is therefore a first parameter which strongly influences the efficiency of an exploration process.

Random graphs

These remarks are confirmed for ER graphs by the grayscale plots in Figures 2 and 3. When the average degree is quite small, there is no qualitative difference between ASP and USP (there exists in general very few shortest path between any two nodes) and the quality of the view is good even for small numbers of sources and destinations.

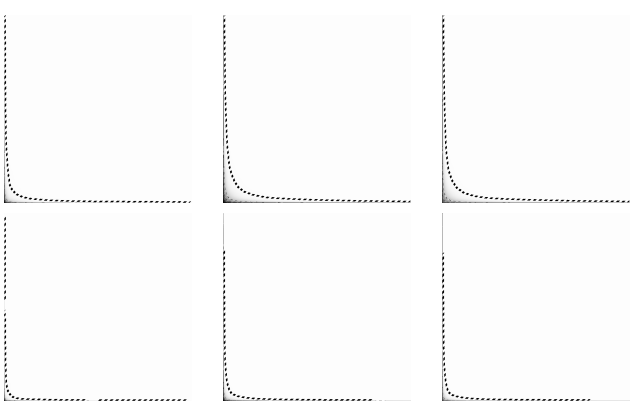

Fig. 2. ER graph: number of nodes, number of links, and average degree. $k = 10$, $N = 10^4$, USP (first row) and ASP (second row).

On the contrary, when the average degree grows, so does the number of shortest paths, and the difference between ASP and USP becomes significant. This can be observed in Figure 3, where we show the plots for both USP and ASP on an ER graph with high average degree. In this case, the nodes are not harder to find than in a low-average degree graph, but the links are.

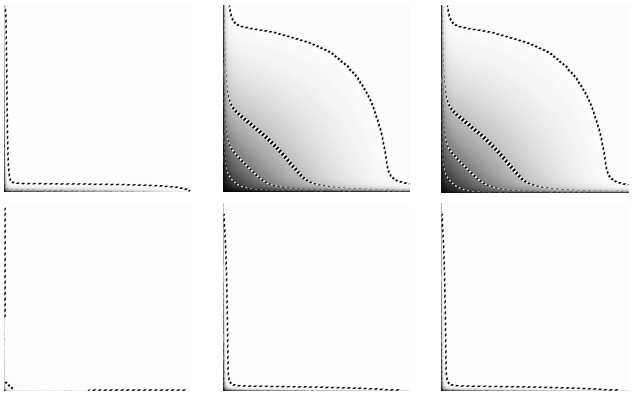

Fig. 3. ER graph: number of nodes, number of links, and average degree. $k = 100$, $N = 10^4$, USP (first row) and ASP (second row).

The fact that the average degree is obtained by dividing two other properties which are improved by the use of more sources and/or destinations has important consequences. If one of the two properties is highly biased and the other is not, then the average degree will have a strong bias. The quotient acts like a *worst case* filter. Figure 3 shows this effect on dense ER graphs. Since the number of links is very poorly estimated, so is the average degree.

In Table II we give a few more precise values extracted from the previous plots, which are of practical interest since the number of sources and destinations are small but greater than those used in current exploration. Indeed, on the Internet, using only 0.1% of nodes as sources means using several thousands sources and only one recent project [53] approaches this nowadays. However, even with this number of sources, an ER graph even with a low average degree cannot be explored in a satisfactory way in the USP case. In order to get a nearly perfect view of the network in terms of links, one has to use at least 1% of the nodes as sources in a network with low average degree.

Still concerning ER graphs, let us observe that, as announced, there is no qualitative difference when one changes the size of the graph: the grayscale plots for a 10^3 nodes ER graph (Figure 4) and the ones for a 10^4 nodes ER graph with the same average degree (Figure 2) are very similar. Notice however that when N grows, the *proportion* of sources and destination necessary to obtain an accurate view decreases, even if the *number* of sources and destinations increases.

		USP		ASP	
src	dest	$k = 10$	$k = 100$	$k = 10$	$k = 100$
0.1%	25%	48%	5.6%	83%	53.5%
0.1%	50%	68.6%	10.5%	94.7%	77.6%
1%	25%	99%	32.2%	99.8%	92.6%
1%	50%	99.9%	54.3%	100%	96.8%

TABLE II

INFLUENCE OF THE AVERAGE DEGREE k AND THE NUMBER OF SOURCES AND DESTINATIONS ON THE PROPORTION OF DISCOVERED LINKS. ER GRAPH, $N = 10^4$.

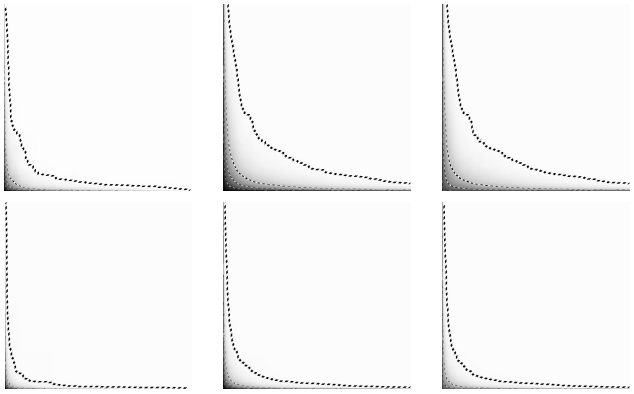

Fig. 4. ER graph: number of nodes, number of links, and average degree. $k = 10$, $N = 10^3$, USP (first row) and ASP (second row).

Scale-free graphs

Let us now observe what happens when we consider scale-free graphs. Let us begin with the AB model which makes it possible to obtain scale-free graphs with a given average degree (by choosing the number of links created for each new node). In Figure 5 (all the plots, using different parameters, display a very similar behavior), we can see that the efficiency of the exploration on such graphs is qualitatively similar to the one on ER graphs, though it is lower. If we want a very precise map, however, we need much more sources and destinations. There is also a strong difference between USP and ASP, which tends to show that there are multiple shortest paths between nodes.

If we make the same experiments with MR graphs using a power law distribution, which also have a scale-free nature and should be equivalent to AB graphs, we obtain the surprising results plotted in Figure 6: the quality of the obtained view is much worse for MR graphs than for AB graphs. Even when considering ASP, one needs to take about half

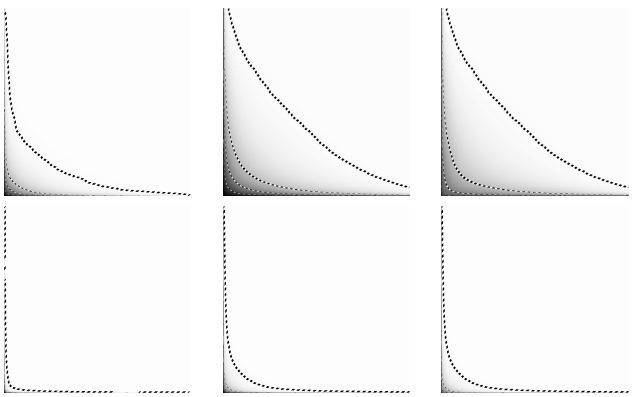

Fig. 5. AB graph: number of nodes, number of links, and average degree. $k = 10$, $N = 10^4$, USP (first row) and ASP (second row).

sources and destinations to view 75% of the graph (both in terms of links and nodes).

Notice also that the average degree is surprisingly well estimated, even if overestimated. Indeed, since the average degree is the quotient of the proportion of nodes and links discovered, if the two properties have the same kind of bias, this may be hidden by the quotient: the evaluation of the average degree is good whenever the ratio between the number of links and the number of nodes is accurate, even if these numbers themselves are wrong. Figure 6 displays such a behavior. Actually the average degree is overestimated since high degree nodes and some of the links attached to them are first discovered and low degree nodes are discovered only in the last steps of the exploration.

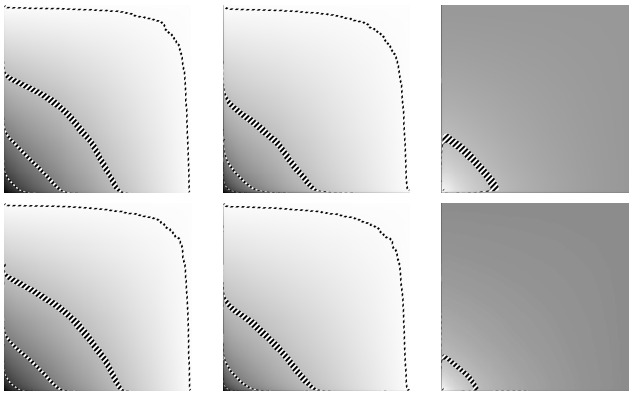

Fig. 6. MR graph: number of nodes, number of links, and average degree. $\alpha = 2.5$, $N = 10^4$, USP (first row) and ASP (second row).

The fact that MR graphs using power law distributions are harder to explore than AB ones rely on a simple explanation: in an AB graph with average degree k, the minimal degree is $\frac{k}{2}$ (we add $\frac{k}{2}$ links at each step, see Section I). On the contrary, in a

MR graph, the number of low-degree nodes (and in particular the number of nodes with only one link) is very high. During an exploration process, these nodes are difficult to discover since they lie on very few shortest paths. For example, a node of degree 1 and the link attached to it are discovered only when we choose this node as a source or a destination. If the number of such nodes is high then the estimation of the size of the graph will be poor.

These explanations can be checked as follows. Instead of considering the original MR graph, we consider its *core* defined as the graph obtained by removing all the nodes of degree 1 and iterating this process until there is no such node anymore. In other words, the graph is composed of the core, to which are attached some tree structures, which we remove. If we run the exploration on the core of a MR graph, we obtain the plots in Figure 7. These results are more in accordance with the ones for the AB graphs. Notice however that it is not only difficult to find a node of degree 1, but also to find all the nodes of low degree, which explains the difference between AB (no nodes of degree lower than $\frac{k}{2}$) and the core of MR graphs.

The difference between ASP and USP is more important in AB graphs than in MR (or in the core of MR), which shows that there are more multiple shortest paths in an AB graph than in a MR one.

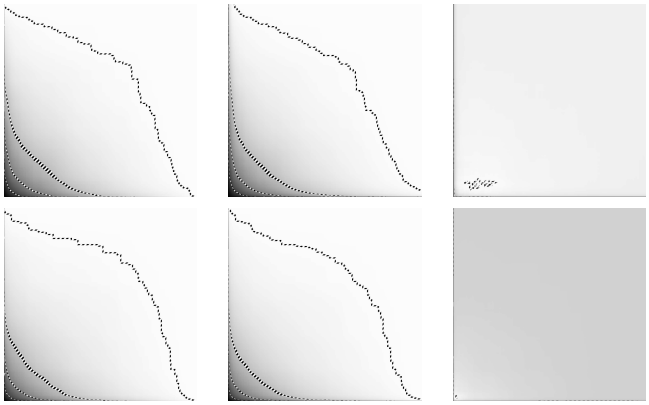

Fig. 7. Core of a MR graph: number of nodes, number of links, and average degree. $\alpha = 2.5$, $N = 10^4$, USP (first row) and ASP (second row).

The important point here is that the quality of an exploration of a MR graph is low because of the large number of low-degree nodes induced by the chosen degree distribution. Such nodes, among which are tree-like structures, are difficult to discover since they lie on few shortest paths, whereas the core of the graph and especially the nodes of high degree are rapidly discovered.

Clusterized graphs

Let us now consider a DM graph, in which there are many triangles and the degree distribution follows a power law. Like in an AB graph, there is no node with only one link. Therefore, the effect noticed above in MR graphs should not appear.

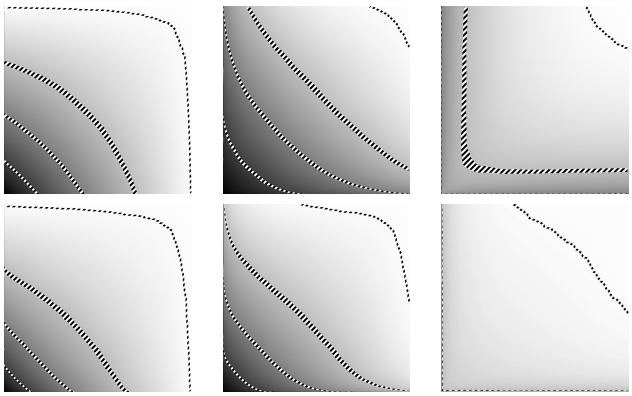

Fig. 8. DM graph: number of nodes, number of links, and average degree. $N = 10^4$, USP (first row) and ASP (second row).

However, one can see in Figure 8 that we again obtain low quality maps of this kind of graphs. The fact that the plots for USP and ASP are very similar indicates that there are very few different shortest paths between nodes. This, and the fact that the quality of the obtained views is low, can be understood as follows. When one wants to explore a clique (complete graph), or more generally a dense graph, one has to use a large number of sources and destinations. For instance in a simple triangle, two links cannot be discovered simultaneously by one traceroute. Therefore three traceroute from wisely chosen sources and destinations have to be processed to discovered a triangle. The same happens for a k-clique in which $k \cdot (k-1)/2$ traceroute have to be processed. The high clustering in DM graphs is equivalent to the fact that there are many subgraphs which are cliques or almost. All these parts of the graph are difficult to explore.

Notice that this time the average degree is poorly estimated, which shows that inferring the average degree is very closely related to the estimation of the number of nodes and links discovered. Very similar behaviors (see Figures 6 and 8 for instance) may lead to very different average degree estimations. This

warns us against drawing fast conclusions concerning properties obtained by dividing a property by another one.

Finally, the conclusion of this section is the following: concerning the number of discovered nodes and links, two properties of graphs make them hard to explore in different ways. The first one is the large number of tree-like structures around the core of the graph. The second one is the high clustering which induces many dense subgraphs. The two properties are complementary and act on different parts of the graph (on the border and on the core, respectively).

III. Average distance

When one uses a few sources and destinations to explore a graph, the obtained view may not be connected. In this case, the average distance does not really make sense. However, the view rapidly becomes connected and we can then estimate the average distance in this view (by computing it exactly for a few random couples, which converges rapidly to the real value).

Notice that, once we have discovered all the nodes, adding new sources and/or destinations decreases the average distance. We therefore begin by overestimating it, and then it converges to the real value. Likewise, when all nodes have been discovered, the USP exploration gives larger values than the ASP one. Therefore, the ASP exploration is more efficient for the evaluation of the average distance. Since the USP exploration is already efficient, we do not display the plots for ASP.

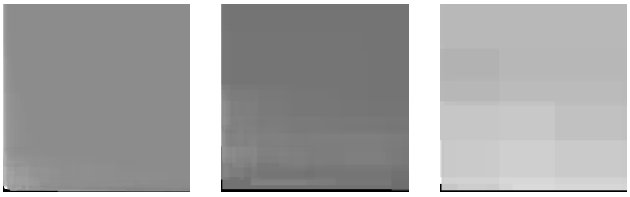

Fig. 9. Average distance for (from left to right): ER graph ($k = 10$, $N = 10^4$), AB graph ($k = 10$, $N = 10^4$), and DM graph ($N = 10^4$). USP.

As one can check in Figure 9, the evaluation of the average distance rapidly becomes very good in all the cases. The plots are nearly uniformly gray, which means that a single `traceroute` is generally a good representative for the average distance in the whole graph. This is a consequence of the fact that distances in a random graph are centered on the average value, see for instance [16], [21]. This is also true, even if the deviation is greater, for *real* internet routes. See [36] and references therein. Results for ER graphs with various average degrees are very similar, and the results for MR graphs are similar to the ones for AB graphs, therefore we do not present these plots here.

Notice also the presence of a black horizontal line at the bottom of the plots which correspond to the fact that the exploration with few destinations yields a set of small graphs (there is no large connected component) which have a very small average distance.

The evaluation of the average distance is slightly less precise for DM graphs (the grey is less uniform). This is due to the fact that clustering induces shortcuts which make it possible to (slowly) reduce the distances when we discover more links. Since the discovery of the links of a DM graph is not very efficient (see Figure 8), the value for the average distance is refined when the number of sources and destinations grows. /

IV. Degree distribution

The degree distribution of the Internet has recently received much attention. It is the main property for which the bias induced by the exploration has been studied [13], [31], [33], [35], [52], [56]. In particular in [13], [35] it is shown that under simple assumptions it is possible to obtain a view with an heterogeneous degree distribution from an ER graph. We will deepen these study here by considering several models, exploration methods, and numbers of sources and destinations.

The question we address here is the following: how fast does the observed degree distribution converge to the real one with respect to the number of sources and destinations? One may use the same kinds of plots as above to answer this question, but this would mean that we need a real-valued test to compare two distributions. Such tests exist (for instance the Student t-test or the Kolmogorov-Smirnov goodness-of-fit test), and such an approach would be relevant here. However, we seek precise insight on *how* the real degree distribution is approached, which makes more relevant the approach consisting in plotting of results for representative values of the parameters. Indeed, these plots make it possible to observe the qualitative difference (e.g. power-law vs Poisson) between the distributions more easily. Finally,

like in the rest of the paper, we conducted extensive simulations and we selected the most relevant ones for this presentation.

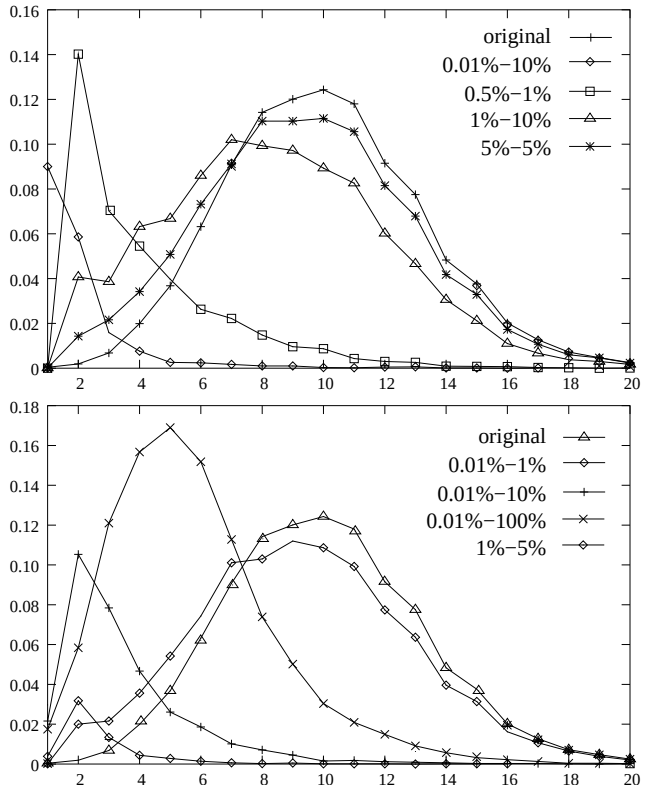

Fig. 10. ER graph: degree distribution. $k = 10$, $N = 10^4$, USP (top) and ASP (bottom).

Let us first consider ER graphs with low average degree. As shown in Figure 10, if the number of sources is very low then the obtained degree distribution is far from the real one. With an USP exploration, the obtained degree distribution converges quite slowly: it is still significantly different from the real one if we take 1% of sources and 10% of destinations. With an ASP exploration, the accuracy is much better: the view is almost perfect even with only 0.5% of sources and 20% of destinations.

The case of ER graphs with high average degree (Figure 11) is more interesting: the presence of high degree nodes makes it possible to obtain heterogeneous degree distributions, well fitted by power laws, with partial USP explorations. This has been studied in previous works [35], [52] to show that the exploration bias may be qualitatively significant. This measurement bias occurs when one considers very few sources and many destinations (Figure 11, top) and the USP exploration. It disappears when one considers a larger number of sources, for instance 0.5% of the whole (Figure 11, bottom), or when one considers an ASP exploration (Figure 12), even for small numbers of sources and destinations.

Notice also that, in intermediary cases, one may obtain surprising results like the plot for 5% of sources and 50% of destinations in Figure 11, which has two peaks. As explained in [35], this is due to the fact that in such cases most of the links close to the sources are discovered, whereas the ones close from the destination are not. The rightmost peak then corresponds to nodes close from the sources (for which we have all their links) while the leftmost one corresponds to the nodes close from the destinations (for which we miss almost every link).

These first results concern ER graphs, for which the degree distributions are not power-laws. They show that it is quite difficult to obtain an accurate view of the degree distribution of such graphs, which is improved significantly by the use of many sources and destinations. As already noticed, the use of a low number of sources may even give degree distributions qualitatively different from the real ones.

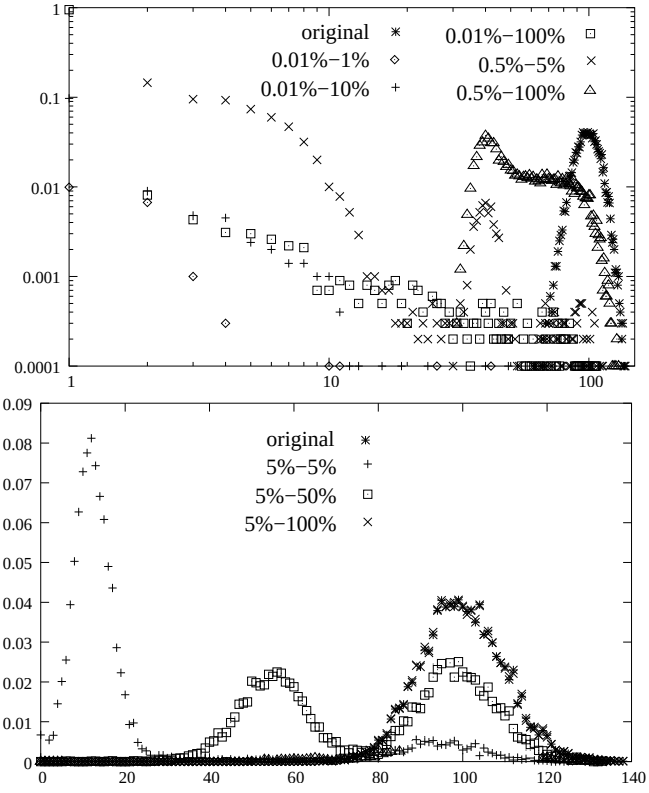

Fig. 11. ER graph: degree distribution. $k = 100$, $N = 10^4$, small number of sources (top log-log scale) and large number of sources (bottom normal scale). USP

If we now consider scale-free graphs, the results are totally different: as one can check in Figures 13 and 14 respectively for MR and DM graphs, both

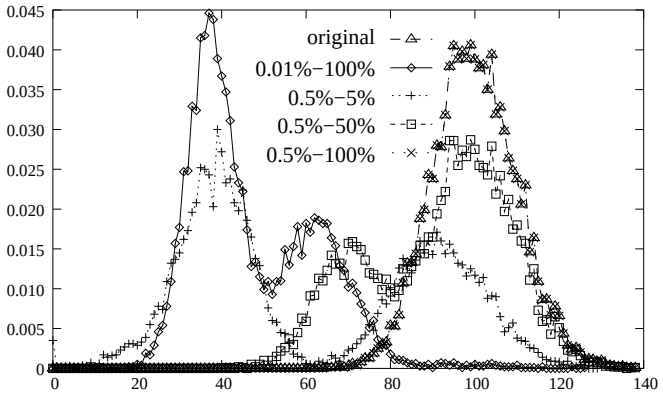

Fig. 12. ER graph: degree distribution. $k = 100$, $N = 10^4$, ASP.

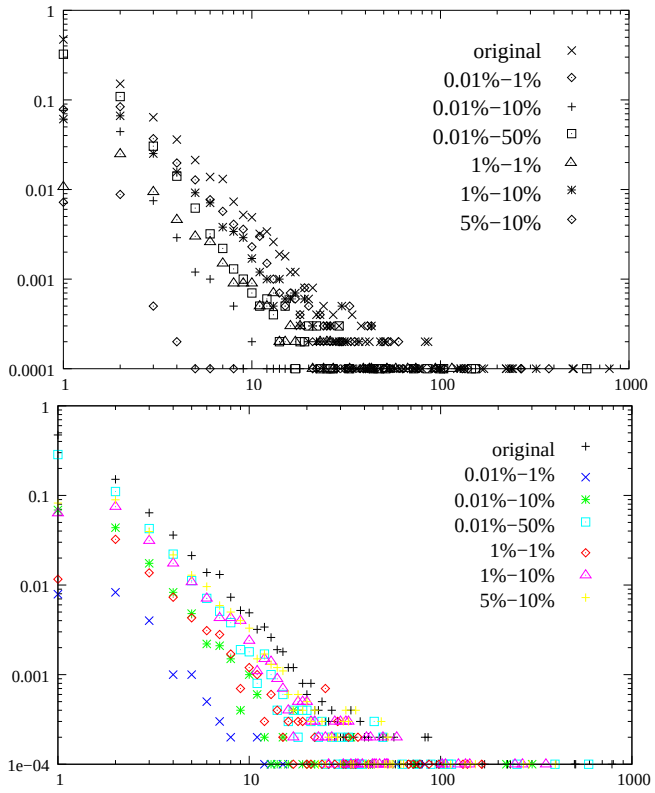

Fig. 13. MR graph: degree distribution. $\alpha = 2.5$, $N = 10^4$, USP (top) and ASP (bottom).

USP and ASP explorations give accurate views of the actual degree distribution [6], even for small numbers of sources and destinations. In the case of MR graphs (the results are similar for AB graphs), the fit is excellent. In the case of DM graphs, the obtained exponent is slightly lower for small numbers of sources

[6]The important characteristic of a power-law distribution is its exponent α, *i.e.* the slope of the log-log plot. Here, we divide the number of nodes of a given degree by the total number of nodes N, including the ones which are not discovered during the exploration in concern. This does not change the slope and makes it possible to plot the distributions in a same figure.

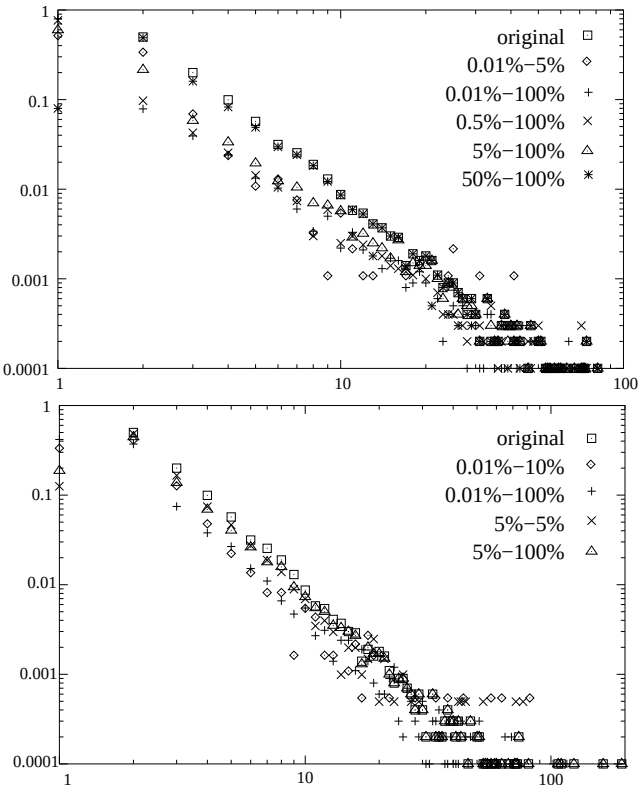

Fig. 14. DM graph: degree distribution. $N = 10^4$, USP (top) and ASP (bottom).

but it rapidly converges to the real one.

In conclusion, the behaviors of ER and scale-free graphs are completely different concerning the accuracy of the obtained degree distributions. Whereas it is quite difficult (especially using an USP exploration) to obtain an accurate estimation for ER graphs, the exponent of the power-law degree distribution of a MR, an AB or a DM graph is correctly measured even with a small number of sources and destinations. Despite the fact that using a very small number of sources and a large number of destinations may give us a wrong idea of the actual degree distribution of a graph, we have shown that these cases are pathological. Indeed, as soon as the number of sources grows, this effect disappears.

V. Clustering

The clustering of a graph is computed by dividing the number of triangles in the graph by the number of connected triples (see Section I). Just like the average degree depends on the obtained numbers of nodes and links (see Section II), this means that the evaluation of the clustering of a graph we obtain using an exploration depends on how fast we discover

triangles with respect to the speed at which we discover triples: the evaluation of the clustering is accurate if we discover a proportion of the total number of triangles similar to the proportion of the total number of triples we discover. We will therefore study how triangles and triples are discovered, together with the clustering itself.

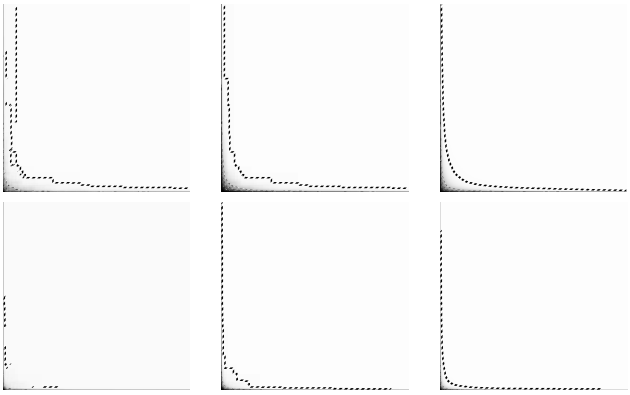

Fig. 15. ER graph: clustering, number of triangles, and number of triples. $k = 10$, $N = 10^4$, USP (first row) and ASP (second row).

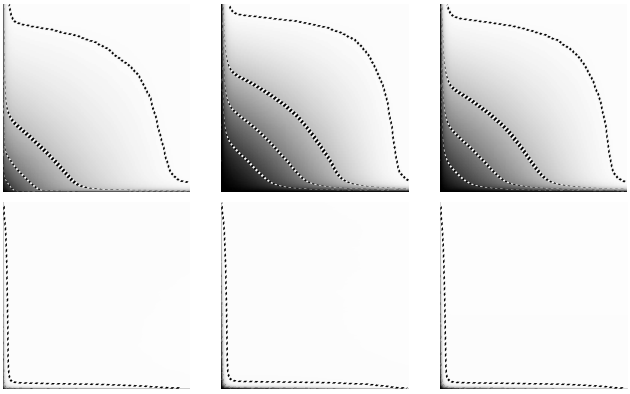

Fig. 16. Dense ER graph: clustering, number of triangles, and number of triples. $k = 100$, $N = 10^4$, USP (first row) and ASP (second row).

Let us first observe what happens for ER graphs. Notice that when the average degree is low, there are almost no triangles in such graphs (and so the clustering is zero). When the average degree grows, so does the clustering. We therefore perform our measurements in both cases. As one can check in Figures 15 and 16, there is no real surprise: increasing the numbers of sources and destinations increases the evaluation of the clustering, a consequence of the fact that the speeds at which triangles and triples are discovered are quite the same. This is in agreement with the results in the previous section which highlighted the fact that dense sub graph are quite hard to explore.

If we turn to AB and MR graphs (the behaviors of

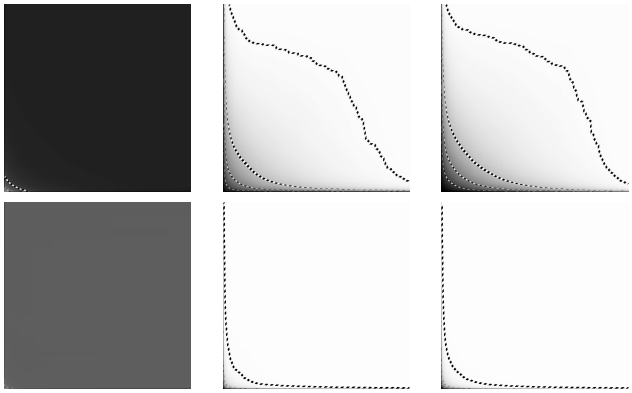

Fig. 17. AB graph: clustering, number of triangles, and number of triples. $k = 10$, $N = 10^4$, USP (first row) and ASP (second row).

the two kinds of graphs are very similar), see Figure 17, we again have a very low clustering but in the USP case it is over-estimated when we consider few sources and destinations. This is a consequence of the fact that we discover much more triangles than triples at the very beginning of the exploration. However, the estimations rapidly becomes accurate, and lower than the initial value. This can be seen in Figure 17: the dark value corresponds to the clustering of the original AB graph, and the only cases where the estimation is wrong are in the lower left corner. The ASP explorations give more accurate results.

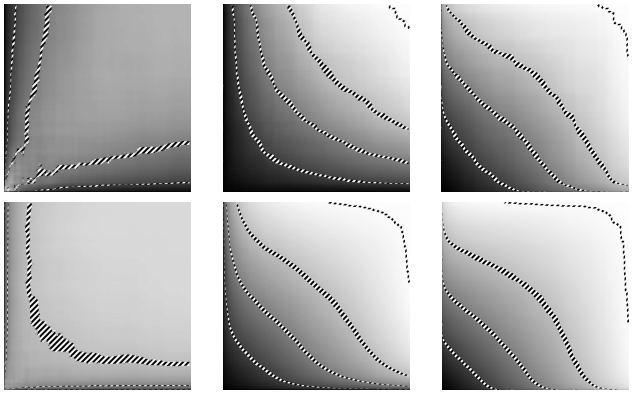

Fig. 18. DM graph: clustering, number of triangles, and number of triples. $N = 10^4$, USP (first row) and ASP (second row).

Let us now observe what happens with a highly clusterized graph, obtained with the DM model. In Figure 18, we can see that the clustering is well evaluated in all the cases, except if we use much more sources than destinations or conversely (notice that this is currently the case for the explorations of the Internet). Indeed, in these cases, there is a strong difference between the speed at which we discover triangles and triples. When the numbers of sources

and destinations are similar, on the contrary, despite we miss many triangles and triples, the proportions we miss of each are similar. In this case, therefore, the estimation of the clustering is accurate.

In conclusion, we see in this section that when we compute an exploration of a graph with low clustering we may over-estimate the clustering. This is due to the fact that the views we obtain are constructed by merging tree-like structures, which makes the triangles hard to discover. It seems however that the obtained evaluation of the clustering is quite accurate even for small number of sources and destinations if the underlying graph has a low clustering. On the contrary, if the graph is highly clusterized, we need both a large number of sources and destinations to obtain a good estimation because discovering triangles is difficult. This is particularly true when the number of sources or the number of destinations is quite low, which implies that the obtained view is a merging of a few trees and therefore over-estimates the number of triples but contains very few triangles.

VI. Sources and destinations placement

Since not all nodes in the Internet play the same role (there are highly connected nodes whereas most have only a few connections, for instance), one may wonder if it is possible to design placement strategies for sources and/or destinations which improve the exploration process. We investigate this idea in this section.

The first well known difference between nodes in the Internet is their degree. We will therefore consider the three following simple strategies in which we choose sources and destinations nodes in an order depending on their degrees. First we can choose both sources and destinations in increasing order of degrees, therefore we will first conduct traceroutes between low-degree nodes. In another strategy, both sources and destinations can be chosen in decreasing order of degree, and finally sources can be chosen in increasing order of degree whereas destinations are chosen in decreasing order. The other strategy (decreasing-increasing) is symmetric. The results should be compared to the ones obtained when we consider sources and destinations chosen at random, as we always do in the rest of the paper. Notice that many other strategies, based on other statistics, are possible. We only present the simple ones based on the degree, which already gives good insight on what may happen. For the same reason, we focus on the basic statistics, namely the proportions of nodes and links discovered, and the average degree.

As one may have expected, these strategies make no real difference on ER graphs. Indeed, in these graphs, all the nodes have almost the same degree. Moreover, the quality of the obtained view is good, even for small numbers of sources and destinations (see Figure 2), therefore one cannot expect to improve it drastically using any strategy. Likewise, the quality of the exploration of AB graphs is already good even for reasonable numbers of sources and destinations. Therefore, even if placement strategies improve the exploration, the difference is not significant.

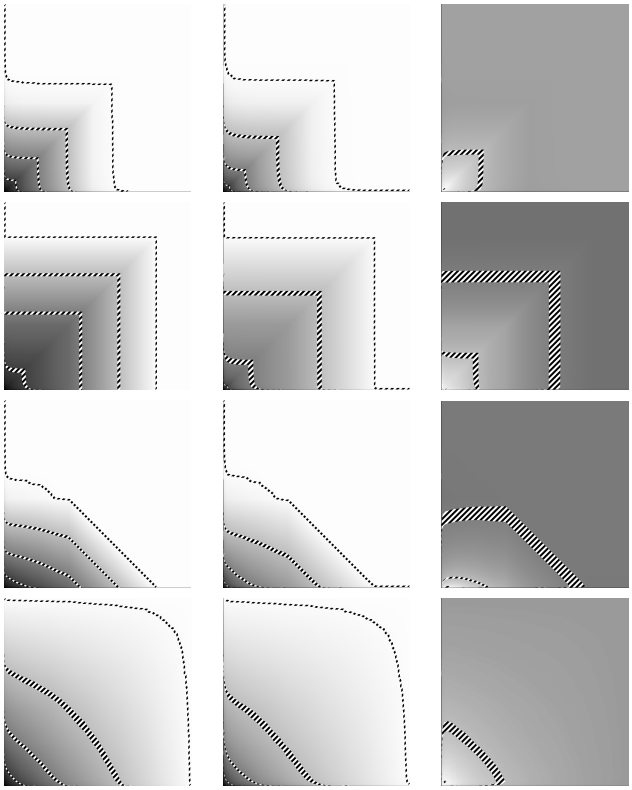

Fig. 19. MR graph: number of nodes, number of links, and average degree. $\alpha = 2.5$, $N = 10^4$, USP with four strategies (from top to bottom): increasing-increasing, decreasing-decreasing, increasing-decreasing, and random. The ASP plots are very similar.

The first case for which the placement strategies are interesting to study is the case of MR graphs, see Figure 19. The obtained results show that sources and destinations placement is definitively relevant: the three strategies give different results, also different from the random strategy. Moreover, the best strategy seems to be the increasing-increasing one. This comes from the fact that, in scale-free graphs,

it is difficult to discover low degree nodes (see Section II), whereas they are taken as sources and destinations in this strategy (and therefore we "discover" them quickly). This explains that the increasing-increasing strategy significantly improves the obtained view, whereas the decreasing-decreasing strategy is inefficient. Notice also that the average degree is overestimated in all cases, even with the increasing-increasing strategy. However with this strategy we ensure the discovery of low degree nodes first and the average degree converges faster to its true value.

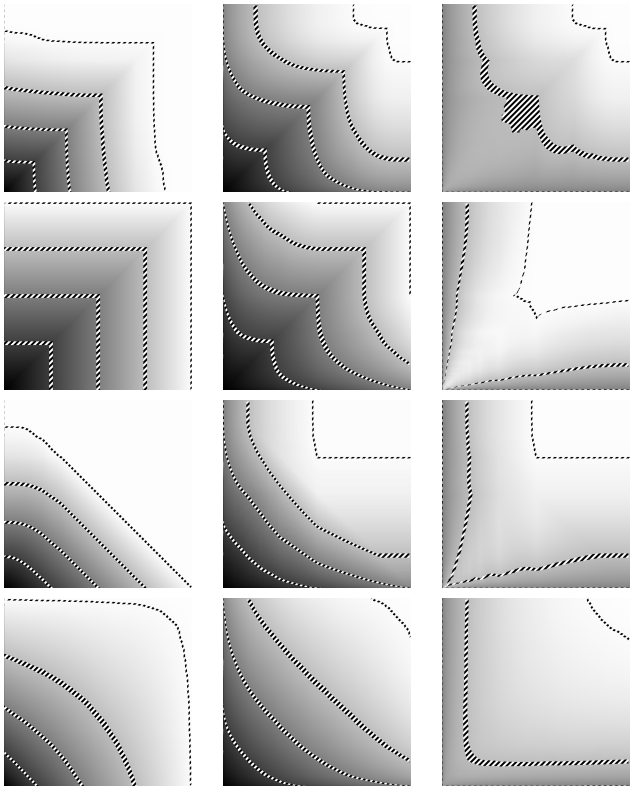

Fig. 20. DM graph: number of nodes, number of links, and average degree. $N = 10^4$, USP with four strategies (from top to bottom): increasing-increasing, decreasing-decreasing, increasing-decreasing, and random. The ASP explorations give very similar results.

Finally, let us observe what happens on DM graphs, see Figure 20 for USP explorations (ASP ones give very similar results). In this case, the best strategy depends on one's aim. If the priority is to discover a large number of nodes using few sources or few destinations, then the best strategy is certainly the increasing-increasing one. This comes from the fact that DM graphs, like MR ones, have a power-law degree distribution and so low-degree nodes are difficult to discover.

However, this strategy gives low performance for the discovery of links, and gives a highly biased average degree. If these properties are of prime interest, one may prefer the increasing-decreasing strategy, which also has the advantage of being efficient if the number of sources and the number of destinations are more or less equal. This strategy actually has very good performance, in particular if we seek a very accurate view of the graph: it significantly improve the number of points above the 0.99-line level. This can be understood as a consequence of the fact that there is a high heterogeneity between sources and destinations.

In conclusion, we see that placement strategies can be used to improve significantly the efficiency of the explorations, but the choice of an appropriate strategy is not trivial. Indeed, it depends both on the properties of the underlying graph and on one's aim. These results are also helpful in understanding the results obtained in previous sections. For instance, they confirm that low degree nodes are difficult to discover, which plays an important role in our ability to map the network.

VII. Real-world data and experiments

Until now, we presented simulations carried out on models of networks and using simple models for `traceroute` and the exploration process. We will now make the same kind of experiments on real-world data to evaluate the relevance of these simulations.

To achieve this, we will use two the following data sets:

- The first one is a well known map of the Internet called *Mercator* [27], [28]. It is obtained by using massively `traceroute` from only one source but with *source routing* and several other improvements. This map has all the properties we have mentioned: high clustering, power-law degree distribution and low average distance. We will focus on the *core* of this graph, *i.e.* the subgraph obtained by iteratively removing the nodes of degree 1. Indeed, we have already seen that the tree-like structures around it are difficult to discover, and our aim is now to identify other properties which may influence the exploration.
- The second data set we will use is the *nec* mapping [44], [42]. It is obtained using 282 sources distributed around the world (public looking-glasses) processing `traceroute` probes from

these sources to 282 destinations chosen at random in a given set of roughly one thousand IP addresses. The number of sources is therefore huge compared to classical explorations (about ten times higher) whereas the number of destinations is quite small. This data set also has an important advantage: we do not only have the map itself but also the actual routes used to construct it. As we will see in the following, it will make it possible to deepen some interesting issues.

A. Comparison with models

Using these two *real-world* graphs, we conducted the same experiments as the ones presented above and we compared the results with the ones obtained on a random graph having exactly the same degree distribution (MR model) and on graphs having the same distribution of clique sizes (GL model). The results for the basic statistics are presented in Figure 21 and in Figure 22 for the core of *Mercator* and for the *nec* graphs, respectively. The results concerning the clustering are plotted in Figure 23 and 24[7]. The results concerning the average distance and the degree distributions are very similar to the ones observed on models, therefore we do not discuss them further.

From Figure 21, Figure 22 and the ones discussed before, we can derive the following interpretations, which are quite similar for the *Mercator* and the *nec* graphs, despite the different ways they have been obtained:

- the difficulty in exploring these graphs is not only due to the presence of tree-like structures around the core, since we removed them in the *Mercator* graph, and since the *nec* graph has almost no tree-like structure,
- these graphs cannot be viewed as MR graphs since the exploration of this kind of graphs gives different results, despite the fact we took MR graphs with the very same degree distributions,
- the clustering could be viewed as the main property responsible for the low quality of the explorations, since the results for the *Mercator* and *nec* graphs are quite similar to the ones for DM graphs (Figure 8, first row) and to the ones for GL graphs (Figures 21 and 22, third rows).

[7]The jumps in the grayscale plots for the clustering of the core of the *Mercator* graph and the *nec* graph are due to the ones in the plot of the number of triples. Themselves are consequences of the fact that, at this point, we take a very high-degree node as a source with many destinations, which suddenly increases the number of triples (by $d(d-1)$ where d is the degree of the node).

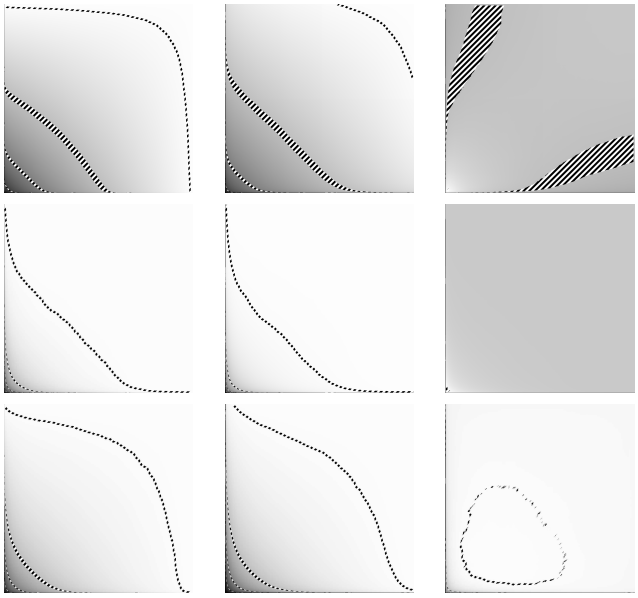

Fig. 21. Number of nodes, number of links, and average degree for (from top to bottom): the core of the original *Mercator* graph, a MR graph with exactly the same degree distribution, and GL graph with the same distribution of cliques sizes. USP explorations.

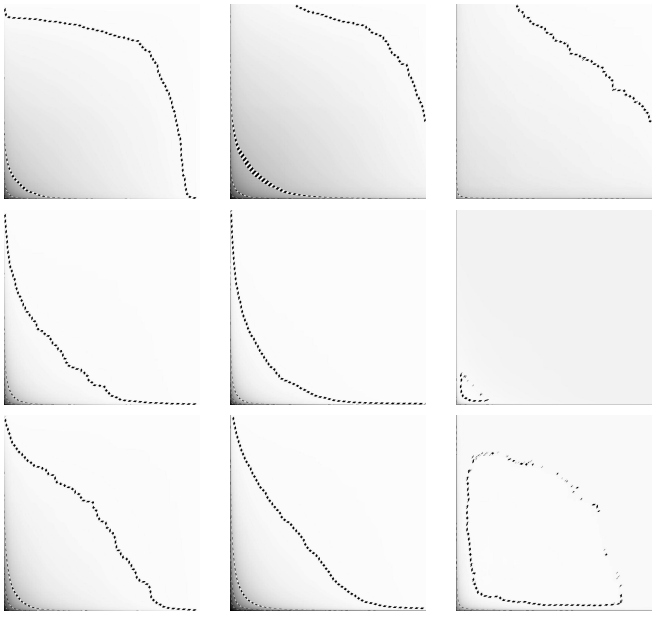

Fig. 22. Number of nodes, number of links, and average degree for (from top to bottom): the *nec* graph, a MR graph with exactly the same degree distribution, and GL graph with the same distribution of cliques sizes. USP explorations.

This last conclusion, however, is not completely satisfactory. Indeed, it appears that no model succeed in capturing really well the behavior or *Mercator* and *nec* graphs concerning the exploration. This

may indicate that other properties than the degree distribution and the clustering may play an important role, see for instance the ones proposed in [39]. This can be checked by observing how the clustering is approximated during explorations of our real-world graphs, and exploration of comparable GL graphs, see Figures 23 and 24. From these figures, it seems that the models do not capture all the properties which influence the exploration process, even if the low degree nodes and the clustering have been clearly identified among them.

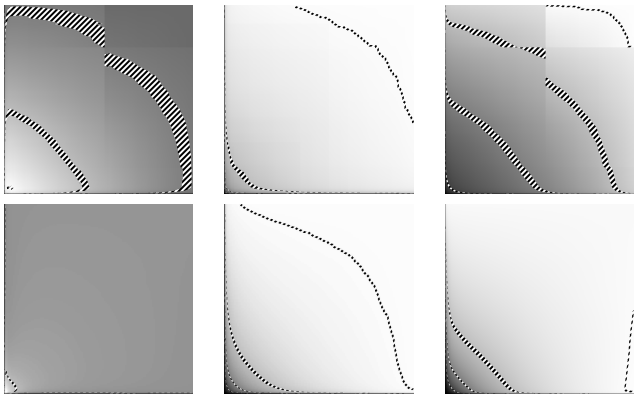

Fig. 23. Clustering, number of triangles, number of triples for the core of the original *Mercator* graph (first row) and a GL graph with the same distribution of cliques sizes (second row). USP explorations.

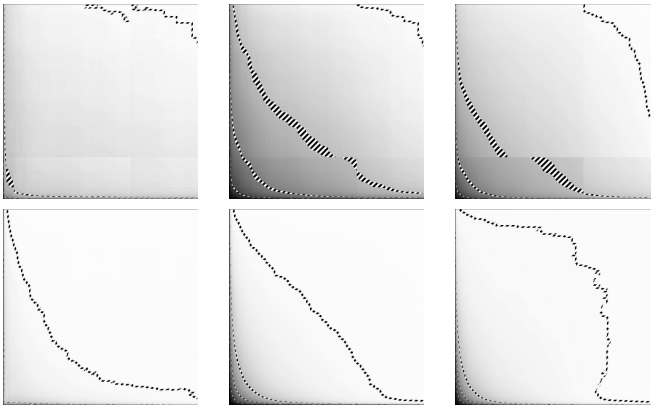

Fig. 24. Clustering, number of triangles, number of triples for the original *nec* graph (first row) and a GL graph with the same distribution of cliques sizes (second row). USP explorations.

B. Going further

The exact sources and destinations, and the obtained routes, used to produce the *Mercator* graph are not available. Therefore, in this case we cannot plot the grayscale plots where we take the same sources and destinations as in the *real* exploration, and where we take real routes rather than shortest paths.

This is possible with graphs obtained using more sources and for which we have the information of which routes have been discovered. We have all this information for the *nec* data set. This makes it possible in this case to compare the grayscale plots obtained using the real `traceroute` paths to the grayscale plots obtained with shortest paths. This is of prime interest since it allows the evaluation of our hypotheses, like for instance the approximation of real routes with shortest paths.

This lead us to compute grayscale plots where we take the same number of sources and destination as in the original exploration (namely 282 each), chosen at random, and in which we approximate routes with shortest paths, just as before (we used both USP and ASP). This gives Figures 25 and 26. Then we compare these plots to the ones obtained when we take the sources and destinations as in the original exploration, and we use the *real* routes discovered by `traceroute`, which gives Figures 27 and 28.

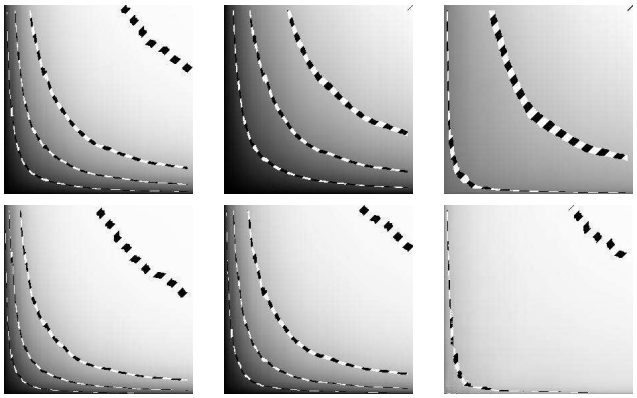

Fig. 25. Number of nodes, number of links, and average degree for the *nec* graph using random sources and destinations and shortest-paths. USP (first row) and ASP (second row).

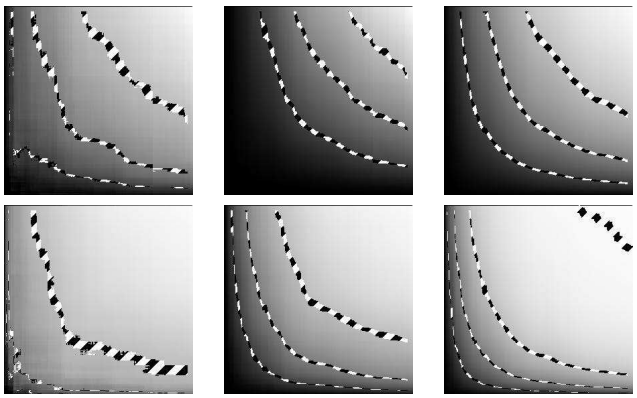

Fig. 26. Clustering, number of triangles and triples for the *nec* graph using random sources and destinations and shortest-paths. USP (first row) and ASP (second row).

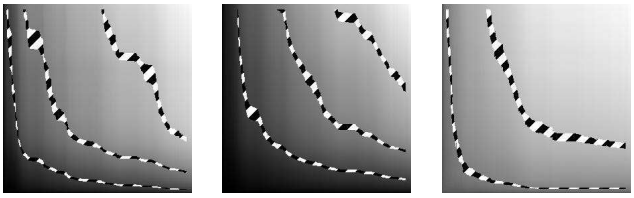

Fig. 27. Number of nodes, number of links, and average degree for the *nec* graph using the *real* routes discovered by `traceroute`.

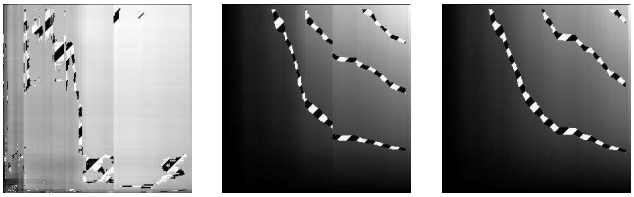

Fig. 28. Clustering, number of triangles and triples for the *nec* graph using the *real* routes discovered by `traceroute`.

The plots fit surprisingly well, the results on the real-world data being in general between an USP and an ASP simulation. This is a very important point, since it gives evidence of the fact that the simulations we conducted throughout the paper rely on reasonable approximations. The results should therefore be considered as relevant, the bias induced by the models of the exploration and of the routes being negligible *from our qualitative point of view*. Let us insist once again, however, on the fact that these results have no meaning from a *quantitative* point of view.

One may also consider the actual number of nodes one obtains using the maximal number of sources and destinations (here, 282), see Table III. When the sources and destinations are the same as in the original exploration, and the routes are the real ones, one sees of course all the graph, and the clustering is the real one. With the models, most nodes are discovered but approximately one quarter of the links are missed. As already explained, this may be a consequence of the presence of links which are between nodes at the same distance from the sources. However, neither an USP nor an ASP exploration can see such links, and Table III shows that here the ASP exploration discovers links much better. Therefore, the poor performance of USP is mainly due here to the fact that there exists several (many) shortest paths between sources and destinations. This indicates that repeating the exploration at several dates may help in improving the maps, since one may then discover several shortest paths.

	nodes	links	cc
original	1.000	1.000	0.087
random nodes/usp	0.997	0.741	0.0079
random nodes/asp	0.999	0.978	0.012

TABLE III

NUMBER OF NODES, NUMBER OF LINKS AND CLUSTERING DISCOVERED WHEN ALL PATHS HAVE BEEN PROCESSED, FOR ORIGINAL ROUTES AND FOR USP AND ASP EXPLORATIONS WITH RANDOM SOURCES AND DESTINATIONS.

CONCLUSION AND DISCUSSION

We conducted an extensive set of simulations aimed at evaluating the quality of current maps of the Internet and the relevance of increasing significantly the number of sources and/or destinations to improve it. To achieve this, we considered the most commonly used models of graphs (namely the ER, the AB, the MR, the DM and the GL ones). Using these simple models has the advantage of making it possible to study separately the influence of various simple statistical properties. We constructed *views* of these graphs and compared them to the original graphs. We focused on the proportion of the graph discovered (both in terms of nodes and links), the average degree, the average distance, the degree distribution and the clustering, which are among the most relevant statistical properties of complex networks in general, and of the Internet in particular.

We presented in this paper our most significant results. To do so, we introduced the grayscale plots and the level lines, which make it possible to give a synthetic view of a huge amount of information, and to interpret it easily. We also discussed how exploration may be improved by placement strategies for the sources and destinations, and we compared the results on network models to the ones obtained on real-world data. This last point confirmed that the simplifications and assumptions we have made in our simulations do not influence significantly the obtained results.

From these experiments, we derive the following conclusions:

- Two statistical properties of graphs influence strongly our ability to obtain accurate views of them using `traceroute`: the presence of many tree-like structures and the high clustering. These two properties act independently and

their effects are combined in the case of the Internet.

- It is relevant to use massively distributed exploration schemes to obtain accurate maps of scale-free clusterized networks like the Internet, in particular if we want to discover most nodes and links, and have an accurate estimation of the clustering. Using more than a few sources should yield much more precise maps.
- On the contrary, the evaluation of the degree distribution of such a network, as well as its average distance, is achieved with very good precision even for reasonably small number of sources and destinations.
- The details of the exploration scheme (for instance USP versus ASP or the behavior of `traceroute`) tends to have little importance when the number of sources and destination grows. In the case of the Internet, this means that distributing explorations can be viewed as a way to improve the independence of the results from the exploration scheme and the details of route properties.
- Despite the fact that power-law degree distribution and high clustering play a role in the efficiency of the explorations of the Internet, it seems that other unidentified properties also influence this efficiency.
- Sources and destinations placement is relevant for the improvement of the explorations, but the choice of the placement is related to the property one wants to capture. Moreover in real measurements, the nodes are indistinguishable before the measurements, therefore such a placement is quite challenging and should be modified during the exploration.

Finally, these results make it possible to conclude that we may be confident in the fact that the Internet graph has a very heterogeneous degree distribution, well approximated by a power law, and that the current evaluation of the exponent of this distribution is quite accurate: current explorations use enough sources to ensure that we do not obtain biased explorations of ER-like graphs, and in the other cases it seems that the estimation of the degree distribution is accurate. Likewise, one might give credit to the available evaluations of the average distance in the Internet. On the contrary, despite the clustering of the Internet is certainly quite high, the estimations we have should be considered more as qualitative than quantitative.

Much could be done to extend our results. First, one may consider more subtle statistical properties, like the correlations between node degrees, or the correlations between degree and clustering. One may also study more precisely some regimes of special interest, like for example the ones currently used (few sources and many destinations), or the one where each source can run `traceroute` a limited number of times. One should also conduct some experiments with more realistic models of `traceroute`. Finally, these simulations results may provide some hints and directions for the formal analysis of the quality of Internet maps. Such studies have began [13], [18], but for now only the degree distribution has been studied in specific cases. Much remains to be done in this challenging direction.

Notice also that we only considered here the *router* level of the Internet and its exploration using `traceroute`. The same kind of study should be conducted at the Autonomous Systems (AS) level and including other techniques like for example the use of BGP tables. The modeling of such techniques is however a problem in itself.

Finally, let us insist on the fact that most real-world complex networks, like the World Wide Web and Peer to Peer systems, but also social or biological networks are generally not directly known. Various exploration schemes are used to infer maps of these networks, which may influence the vision we obtain. The metrology of complex networks is therefore a general scientific challenge, for which the goal is to be able to deduce properties of the real network from the observed ones. The methodology we developed here may be applied to these different cases with benefit.

Acknowledgments. We thank Ramesh Govindan for providing useful data. We also thank Aaron Clauset, Mark Crovella, Benoit Donnet, Timur Friedman and all the `Traceroute@Home` [55] staff for their helpful comments. This work is supported in part by the ACI *Sécurité et Informatique* project *MetroSec* [54].

REFERENCES

[1] D. Achlioptas, A. Clauset, D. Kempe, and C. Moore. On the bias of traceroute sampling, or: Why almost every network looks like it has a power law. In *ACM Symposium on Theory of Computing (STOC2005)*, 2005.

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

[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] Paul Barford, Azer Bestavros, John Byers, and Mark Crovella. On the marginal utility of network topology measurements. In *ACM SIGCOMM Internet Measurement Workshop 2001*, San Francisco, CA, November 2001. ACM SIGCOMM.

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

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

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

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

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

[11] F. Chung and L. Lu. The average distances in random graphs with given expected degrees, 2002.

[12] K. Claffy, T. Monk, and D. McRobb. Internet tomography. Nature Magazine, Web Matters. http: //helix.nature.com/webmatters/tomog/tomog.html.

[13] A. Clauset and C. Moore. Accuracy and scaling phenomena in internet mapping. *Phys. Rev. Lett.*, 2005.

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

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

[16] R. Cohen and S. Havlin. Scale-free networks are ultrasmall. *Phys. Rev. Lett.*, 90(058701), 2003.

[17] L. Dall'Asta, J.I. Alvarez-Hamelin, A. Barrat, A. Vzquez, and A. Vespignan. A statistical approach to the traceroute-like exploration of networks: theory and simulations. In *Workshop on Combinatorial and Algorithmic Aspects of Networking*, 2004.

[18] L. Dall'Asta, J.I. Alvarez-Hamelin, A. Barrat, A. Vzquez, and A. Vespignan. A statistical approach to the traceroute-like exploration of networks: theory and simulations. *Special issue of Theoretical Computer Science on Complex Networks*, 2005.

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

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

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

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

[23] A. Fabrikant, E. Koutsoupias, and C.H. Papadimitriou. Heuristically optimized trade-offs: A new paradigm for power laws in the internet. In *ICALP*, 2002.

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

[25] Cooperative Association for Internet Data Analysis. http://www.caida.org/.

[26] Cooperative Association for Internet Data Analysis Skitter tool. http://www.caida.org/tools/measurement/skitter/.

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

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

[29] J.-L. Guillaume and M. Latapy. Bipartite graphs as models of complex networks. In *Workshop on Combinatorial and Algorithmic Aspects of Networking (CAAN)*, 2004.

[30] J.-L. Guillaume and M. Latapy. Bipartite structure of all complex networks. *Information Processing Letters*, 90(5):215–221, 2004.

[31] J.-L. Guillaume and M. Latapy. Complex network metrology. *Complex systems*, 2005.

[32] J.-L. Guillaume and M. Latapy. Relevance of massively distributed explorations of the internet topology: Simulation results. In *IEEE Infocom 2005*, 2005.

[33] Y. Hyun, A. Broido, and K. Claffy. Traceroute and BGP AS path incongruities. http://www.caida.org/outreach/papers/2003/ASP/.

[34] C. Jin, Q. Chen, and S. Jamin. Inet: Internet topology generator. Technical Report CSE-TR-443-00, Department of EECS, University of Michigan, 2000.

[35] A. Lakhina, J. Byers, M. Crovella, and P. Xie. Sampling biases in IP topology measurements. In *IEEE INFOCOM*, 2003.

[36] J. Leguay, M. Latapy, T. Friedman, and K. Salamatian. Describing and simulating internet routes. In *Networking*, 2005.

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

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

[39] D. Magoni and J.-J. Pansiot. Analysis of the autonomous system network topology. *ACM SIGCOMM Computer Communication Review*, 31(3):26 – 37, July 2001.

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

[41] D. Magoni and J.-J. Pansiot. Influence of network topology on protocol simulation. In *ICN'01 - 1st IEEE International Conference on Networking*, volume Lecture Notes in Computer Science 2093, pages 762–770, July 9-13, 2001.

[42] Damien Magoni. nec (*ne*twork *c*artographer) – https://dpt-info.u-strasbg.fr/∼magoni/nec/.

[43] Damien Magoni. Tearing down the internet. *IEEE Journal on Selected Areas in Communications*, 21:949–960, 2003.

[44] Damien Magoni and Mickaël Hoerdt. Internet core topology mapping and analysis. *Computer Communications*, 28:494–506, 2005.

[45] A. Medina, I. Matta, and J. Byers. On the origin of power laws in internet topologies. In *ACM Computer Communication Review, 30(2)*, pages 18–28, april, 2000.

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

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

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

[49] I. Norros and H. Reittu. On the power-law random graph model of massive data networks. *Source Performance Evaluation archive*, 55(1-2):3–23, 2004.

[50] J. Pansiot and D. Grad. On routes and multicast trees in the internet. *ACM Computer Communication Review*, 28(1):41–50, 1998.

[51] R. Pastor-Satorras and A. Vespignani. *Evolution and Structure of the Internet: A Statistical Physics Approach*. Cambridge University Press, 2004.

[52] T. Petermann and P. De Los Rios. Exploration of scale-free networks. *To appear in Eur. Phys. J. B*, 2004.

[53] DIMES@home Project. http://netdimes.org/.

[54] MetroSec project. http://www.laas.fr/METROSEC/.

[55] Traceroute@Home project. University of Paris 6, coordinator: Timur friedman. http://www.tracerouteathome.net.

[56] P. De Los Rios. Exploration bias of complex networks. In *Proceedings of the 7th Conference on Statistical and Computational Physics Granada*, 2002.

[57] N. Spring, R. Mahajan, and D. Wetherall. Measuring ISP topologies with rocketfuel. In *Proceedings of ACM/SIGCOMM '02*, August 2002.

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

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

[60] R. van der Hofstad, G. Hooghiemstra, and D. Znamenski. Distances in random graphs with finite mean and infinite variance degrees.

[61] R. van der Hofstad, G. Hooghiemstra, and D. Znamenski. Random graphs with arbitrary i.i.d. degrees.

[62] A. Vazquez, R. Pastor-Satorras, and A. Vespignani. Internet topology at the router and autonomous system level. [cond-mat/0206084].

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

[64] E.W. Zegura, K.L. Calvert, and M.J. Donahoo. A quantitative comparison of graph-based models for Internet topology. *IEEE/ACM Transactions on Networking*, 5(6):770–783, 1997.