Impact of Sources and Destinations on the Observed Properties of the Internet Topology

Frédéric Ouédraogo[a,b,c], Clémence Magnien[a,b,*]

[a] *UPMC Univ Paris 06, UMR 7606, LIP6, F-75016, Paris, France*
[b] *CNRS, UMR 7606, LIP6, F-75016, Paris, France*
[c] *University of Ouagadougou, LTIC, Ouagadougou, Burkina Faso*

Abstract

Maps of the internet topology are generally obtained by measuring the routes from a given set of sources to a given set of destinations (with tools such as `traceroute`). It has been shown that this approach misses some links and nodes. Worse, in some cases it can induce a *bias* in the obtained data, i.e. the properties of the obtained maps are significantly different from those of the real topology. In order to reduce this bias, the general approach consists in increasing the number of sources. Some works have studied the relevance of this approach. Most of them have used theoretical results, or simulations on network models. Some papers have used real data obtained from actual measurement procedures to evaluate the importance of the *number* of sources and destinations, but no work to our knowledge has studied extensively the importance of the *choice* of sources or destinations. Here, we use real data from internet topology measurements to study this question: by comparing partial measurements to our complete data, we can evaluate the impact of adding sources or destinations on the observed properties.

We show that the number of sources and destinations used plays a role in the observed properties, but that their choice, and not only their number, also has a strong influence on the observations. We then study common statistics used to describe the internet topology, and show that they behave differently: some can be trusted once the number of sources and destinations are not too small, while others are difficult to evaluate.

1. Introduction

The mapping of the internet topology has received a lot of interest from the research community recently. Obtaining accurate maps of this topology is indeed of key importance for several applications, including protocol simulations.

*Corresponding author
Email addresses: `frederic.ouedraogo@univ-ouaga.bf` (Frédéric Ouédraogo), `clemence.magnien@lip6.fr` (Clémence Magnien)

No exact map is available due to the internet's distributed construction and administration, and obtaining one through measurements is a challenging task.

Efforts have been made to discover the topology at the router or IP level by using tools such as `traceroute`: one collects routes from a given set of sources to a given set of destinations, and merges them to obtain a view of the topology. Such views give much information on the global shape of the internet. It has been shown that the internet topology has some statistical properties which make it very different from most models used previously. This induced an intense activity in the acquisition of such maps, see for instance [25, 3, 24, 12] and in their analysis, see for instance [11, 4, 12, 20, 26].

It must be clear that the image obtained in such a way is partial (some nodes and links are not seen) and may be biased by the exploration process. Several experimental and formal studies have been conducted to evaluate the accuracy of the obtained maps of the internet, as well as the benefit of using more than one source (both concerning the quantity of data gathered and the bias reduction) [1, 2, 5, 7, 9, 10, 13, 14, 15, 16, 17, 21, 23, 26]. All these studies give good arguments of the fact that maps of the internet collected from a single source are very incomplete, and that there probably is a bias induced by the exploration process. This bias is greatly reduced when using more sources for the measurement.

Using real data [18, 22] coming from recent measurements, we study in depth the differences between the IP-level maps that can be obtained from different sources and destinations. Following a methodology introduced in [14, 15], we study the impact of the number of sources and destinations on a number of classical graph properties, which allows us to determine to which extent these properties are biased by the exploration process. We thus confirm results previously established on models. We also study the impact of the choice of sources and destinations on the quality of the obtained view.

We first present the data set we use and our methodology in Section 2. We then present in Section 3 how the choice of sources and destinations, and not only their number, can influence the obtained view and its statistical properties. Section 4 presents our detailed analysis of the impact of the number and choice of sources and destinations on the obtained view of the IP-level topology, by studying classical graph properties. We discuss related work in Section 5, and present our conclusions in Section 6.

2. Data set and methodology

The data we use was collected in [18] and is publicly available [22]. Measurements were conducted from more than 150 monitors. To each monitor was associated a *destination set* that stayed the same for the whole duration of the measurements. The measurements then consisted in periodically running the `tracetree` tool, which collects the routes from a given monitor to a set of destinations in a `traceroute`-like manner, but much more quickly, and imposing a far smaller load on the network. The measurements were conducted with a

high frequency (typically leading to about one hundred measurement rounds per day), for a long period of time (from weeks to several months, depending on the monitor). For a more comprehensive description of the measurement procedure and the obtained data, see [18].

In this original data set, all sources do not use the same destination set. To study the impact of adding sources and destinations, we need to have a set of sources running measurements towards the same destinations. In this paper, we therefore use a data set consisting of a set S of 11 sources, which are associated to the same set D of 3 000 destinations.

For each source j and destination k, we call $g_{j,k}$ the graph composed of the union of paths from source j to destination k [1] (the union keeps only one copy of each link; all graphs we consider have no loops or multiple edges; moreover, all links are undirected and unweighted). The nodes are the IP addresses observed on the paths from source j to destination k, and the links represent the hops at the IP level from node to node. The part of the IP-level topology observed from source j is then the union of what was seen from this source towards all destinations:

$$g_j = \bigcup_k g_{j,k}.$$

Notice that, in each measurement round, if a machine did not reply to a probe, it is represented as a star ($*$). In a given measurement round, all stars are different. A problem arises when we perform the union of several measurement rounds: we cannot know if stars appearing at the same location at different rounds correspond to a same machine or not. The solution we have chosen consists of not taking into account the stars or the links associated to them in the original data set. This causes the paths from the source to some destinations to be disconnected, and therefore the graphs $g_{j,k}$ may not be connected. However, this is not a problem for our purpose, because the graphs g_j are almost connected: in all cases, the largest connected component contains at least 92% of nodes.

All sources are not equivalent, and the sizes of the graphs g_j vary. The smallest has 16 469 nodes, and the largest has 26 447 nodes. More details about this can be found in Section 3.

From this we define the graph G to be the topology observed with all sources: $G = \bigcup_j g_j$. It has $n = 42\,141$ nodes and $m = 165\,438$ links. Note that the nodes of G are IP addresses and not routers, as we do not perform alias resolution.

Our approach consists generally in studying views of G obtained by exploring it using only subsets of the sources and destinations. The important point in this study is that our data provides us with the *actual* paths from sources to destinations. We therefore do not rely on modeling to obtain the partial views.

[1] i.e. the union of the nodes and links appearing on paths from j to k in all measurement rounds. Because of load balancing and other phenomena, this path is generally not constant throughout the whole measurement period.

If we denote by S' a subset of sources and by D' a subset of destinations, $G_{S'D'}$ is the view of G obtained with theses sources and destinations, defined by:

$$G_{S'D'} = \bigcup_{\substack{j \in S' \\ k \in D'}} g_{j,k}.$$

In the rest of the paper, we will therefore evaluate the impact of the choice of sources and destinations by comparing different views $G_{S'D'}$, for different subsets S' and D' of the sources and destinations, with different sizes.

3. Impact of the choice of sources and destinations

In this section we evaluate the impact of both the choice and the number of sources and destinations on the obtained view. We focus first on the number of nodes and links of the obtained view, then on more elaborate properties.

3.1. Number of nodes and links

We focus here on the observed number of nodes and links. We will see that the choice of sources and destinations has a strong impact on the observations.

Studying a large number of views obtained with randomly chosen sets of sources and destinations is computationally expensive. To solve this problem we use the following method. When studying the impact of the choice and number of sources (resp. destinations), we will always build a view obtained by k sources and all destinations (resp. k destinations and all sources) by adding one source (resp. destination) to a view obtained with $k-1$ sources and all destinations (resp. $k-1$ destinations and all sources). Starting with a single source (resp. destination) and adding sources (resp. destinations) one by one allows us to study the impact of the number of sources (resp. destinations). Changing the order in which we consider sources (resp. destinations) allows us to study the impact of the choice of sources (resp. destinations).

A good way of evaluating the impact of the choice of sources is to investigate, given a number k of sources, what are the largest and smallest sizes (in terms of the number of nodes) of the graphs it is possible to obtain by considering k sources. We denote by $M_s(k)$ the maximum size of a graph we can obtain with k sources:

$$M_s(k) = \max_{S' \subseteq S, |S'|=k} |G_{S'D}|,$$

and by $m_s(k)$ the minimum size of such a graph:

$$m_s(k) = \min_{S' \subseteq S, |S'|=k} |G_{S'D}|.$$

We define dually $M_d(k)$ and $m_d(k)$, which are the maximum and minimum sizes of graphs that can be obtained with k destinations.

Obtaining the maximum or minimum function is too computationally expensive. We therefore propose a greedy heuristic for approximating it: at each

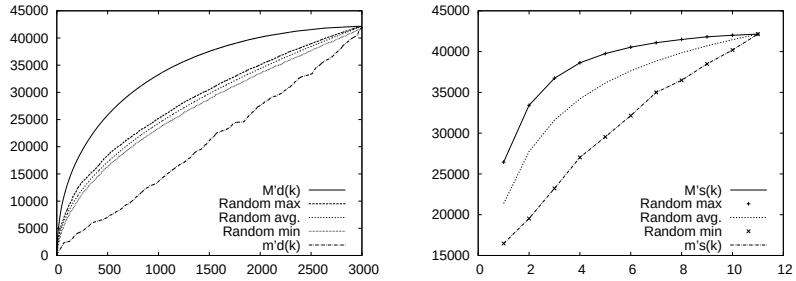

Figure 1: Left: Impact of the choice of destinations on the observed number of nodes. Right: Impact of the choice of sources on the number of nodes of the obtained view.

step we consider the graph obtained in the previous step, and choose the source (resp. destination) that adds the most nodes to this graph. We denote the size of the graph obtained at the k-th step by $M'_s(k)$ (resp. $M'_d(k)$). Conversely, we approximate the minimum by starting with the source (resp. destination) that discovers the fewest nodes, and choose at each time step the source (resp. destination) that adds the least nodes to the current graph. We denote the size of the graph obtained at the k-th step by $m'_s(k)$ (resp. $m'_d(k)$).

We have no guarantee of how close $M'_s(k)$ and $m'_s(k)$ are to $M_s(k)$ and $m_s(k)$, but we know that they are lower and upper bounds for them, respectively.

In order to get an intuition about what we obtain when we select sources or destinations at random, we also computed the number of nodes and links seen with 1 000 random orders. For these orders, we computed for each number k of sources (resp. destinations) the maximum, the minimum and the average value observed with the first k sources (resp. destinations).

The behaviors observed for the number of discovered nodes and links were very similar, therefore we only present the figures concerning the number of nodes.

Figure 1 (left) presents the impact of the chosen destinations on the observed number of nodes. We observe several things. First, there is a high difference between the approximated maximum and minimum sizes $M'_d(k)$ and $m'_d(k)$. For 500 destinations for instance, the observed number of nodes varies from approximately 7 500 to more than 25 000. This shows that, for a same number of destinations, the choice of these destinations may have a dramatic influence on the observed topology.

This difference is not so clear however when we consider *random* orders: the plots for the average, the minimum and the maximum among 1 000 orders are both close to each other and far from the estimated maximum and minimum orders. This seems to indicate that, concerning the number of nodes, there are some atypical orders yielding very different results from the average, but that most orders are close to the average and that it is therefore representative.

Figure 1 (right) shows the impact of the chosen sources on the observed

number of nodes. As for the destinations, the difference between the minimum and the maximum is important: when considering 4 sources, the number of observed nodes varies from 64% to 91% of the whole graph.

However, we can see that the maximum value observed for 1 000 random orders is equal to $M'_d(k)$; in the same way, the minimum value observed is equal to $m'_d(k)$. The fact that we consider a small number of sources plays an important role in this. Though the number of orders we consider is very small compared to the total number of possible orders on 11 sources (1 000 compared to $11! \sim 40 \cdot 10^6$), the probability of observing the actual maximum for small numbers of sources is very high. For $k = 3$ sources for instance, the probability of *not* observing $M_s(3)$ is equal to 0.2% approximately [2].

Our observations are in accordance with previous work [5] about the impact of the number of sources and destinations on the observed number of nodes and links. However, the authors had considered the greedy maximum order for sources, and a random order for destinations. Their conclusions were that there is a diminishing returns effect concerning sources: adding sources provides less and less additional information. Adding destinations, on the other hand, gives an approximately constant benefit.

The authors of [26] have mitigated the diminishing returns effect. They argue that, even though adding sources brings less additional information on average, the benefit of adding many sources is far from negligible. They considered a large number of sources, and sorted them by decreasing order of the number of links they discover. Notice however that this order is naturally close to the greedy maximum order.

By comparing the greedy maximum and minimum orders as well as random orders, both for sources and destinations, we conclude that the effect of adding sources and destinations is similar. We observe a strong diminishing returns effect for the greedy maximum orders: in this case, the last sources or destinations bring very little new information. This effect does not appear for the greedy minimum orders, for which the plots are approximately linear. The question of whether we would observe a diminishing returns effect for the greedy minimum order if more sources or destinations were used remains open, but we conjecture that the linear shape of the plot is caused by an intrinsic heterogeneity in the number of nodes each source or destination discovers. Finally, the average orders represent what one may expect to observe in practice. The diminishing returns effect is present, but much less striking than for the greedy maximum orders. In particular, while the first sources or destinations discover more nodes that the others, adding more sources or destinations still brings an approximately linear benefit. Extreme behaviors, leading to the observation of much more (or much less) nodes than expected, happen much more often for sources than destinations. This is linked in part to the fact that we use less

[2] The probability for a given order to not select the three sources that give $M_s(3)$ is $p = 1 - 3!/(11 * 10 * 9)$, and $p^{1000} \sim 0.2\%$. The same probability applies when considering 8 sources.

Average degree	# sources	# dest.	Global Clust.	# sources	# dest.
7.852343	11	2993	0.160401	1	244
7.856401	11	2976	0.256501	1	5
7.855130	11	2975	0.125881	3	358
7.853955	11	2914	0.496428	1	353
7.852128	11	2997	0.152805	5	28
7.854578	11	2976	0.250000	1	1
7.854771	11	2984	0.154679	2	244
7.891898	11	2934	0.131765	4	396
7.853233	11	2973	0.185765	2	3
7.853207	11	1996	0.136787	3	503
Original graph G					
7.851641	11	3 000	0.101155	11	3 000

Table 1: Maximum values of the average degree and the global clustering reached with 10 different orders on sources and destinations. For each order, we report the maximal values observed over all graphs $G_{S'D'}$ obtained with the first $|S'|$ sources and $|D'|$ destinations in the order.

sources than destinations.

Finally, we conclude that the number of sources and destinations *as well as their choice* plays an important role on the size of the observed topology.

3.2. Other properties

We now study the impact of the choice of sources and destinations on other properties. Given an order on sources and one on destinations, we build all graphs $G_{S'D'}$ obtained with the first $|S'|$ sources and $|D'|$ destinations in the orders. We then compare the observations for different orders.

Table 1 shows the maximum values observed for the average degree [3] and the global clustering [4] for 10 different random orders on sources and destinations. We observe that the maximum is different for all orders. In the case of the global clustering, the difference can be quite high: the observed values vary from 0.13 to 0.5. Moreover, the number of sources and destinations for which it is reached varies also. This means that, for a given number of sources and destinations, their actual choice can have a strong impact on the observed properties. Though the maximal value is not representative of all that can be observed for a given order, this already brings to light strong differences in the observations for different orders.

Notice that the average degree presents much less variation: the maximal values observed are all close to each other, and are obtained for similar numbers

[3] $d^\circ = 2m/n$, see Section 4.1.

[4] gc is equal to three times the number of triangles, divided by the number of connected triples, see Section 4.2.

of sources and destinations. This shows that not all properties are affected in the same way by the choice of sources and destinations, and that some properties can probably be trusted more than others.

4. Grayscale plots

We now turn to a more detailed study of the impact of the choice of sources and destinations on the observed properties. To study this impact on a given property [5] p, we use *grayscale plots* as introduced in [14, 15]. We consider a rectangle of width $|D|$ and height $|S|$. Given an order on sources and one on destinations, each point (d, s) of the rectangle corresponds to the graph $G_{S'D'}$ such that S' contains the first s sources in the order, and D' contains the first d destinations. The point is drawn using a grayscale representing the value of p: from black for $p = 0$ to white for the maximal observed value of p (which might be greater than the value obtained for the whole graph G). The points darker than the upper-right point correspond to conditions where the value p is underestimated, whereas lighter points correspond to conditions in which it is over-estimated. The gray variation is linear: if a dot is twice as dark as another dot, then the associated value is twice as large.

Finally, to improve readability, we represent level lines. 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 $l + 0.01$. The 0.1-level line is represented in white, the 0.5-level line alternates black and white segments, and the 0.9-level line is represented in black [6].

The fact that the graph corresponding to the point $(d-1, s-1)$ is included in the one corresponding to (d, s) quickens considerably the computations needed to produce such plots.

Since, as already discussed, the choice of sources and destinations has an influence on the observed properties, different orders will produce different grayscale plots. Generating plots for different orders, as well as the average plot obtained with several orders, will therefore help us in evaluating this impact.

We now turn to the study of important graph statistical properties. For each property, we will recall its definition, then study grayscale plots corresponding to different orders to evaluate the impact of the number of sources and destinations (and the impact of their choice) on the observations.

4.1. Average degree and density

The degree $d^\circ(v)$ of a node v is its number of links, or equivalently, its number of neighbors. The average degree d° of a graph is the average of the

[5] All properties we consider in this paper are real-valued and non-negative.

[6] Notice that on some plots these lines do not appear, because the variation of the value of the observed property is not large enough.

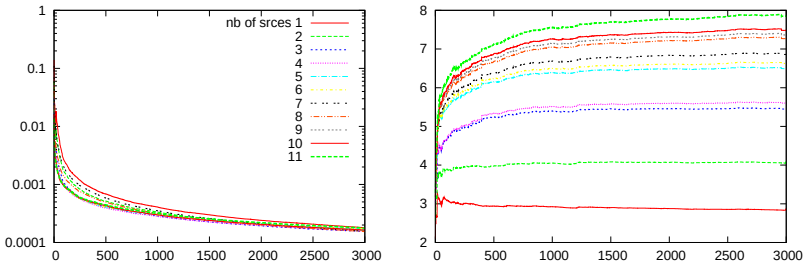

Figure 2: Density (left) and average degree (right) as a function of the number of destinations considered (the order on destinations is a random order). Each plot corresponds to a different number of sources considered. The keys for both plots are the same.

degree over all nodes:

$$d^\circ = \frac{1}{n}\sum_v d^\circ(v) = \frac{2m}{n}.$$

The density is the number of links in the graph divided by the total number of possible links:

$$\delta = \frac{2m}{n(n-1)}.$$

The density indicates up to which extent the graph is fully connected (all links exist). Equivalently, it gives the probability that two randomly chosen nodes are linked in the graph. There is a trivial relation between the average degree and the density: $\delta = \frac{d^\circ}{(n-1)}$.

This relation implies that, when the average degree is constant with respect to the graph size, the density tends to 0 when n grows.

Figure 2 (left) presents the density as a function of the number of destinations considered, for different number of sources. The order chosen for the destinations is a random order. For small numbers of destinations, the number of sources plays an important role: for 100 destinations for instance, the density varies by a factor of approximately 5, depending on the number of sources. When the number of destination grows, however, the difference quickly becomes very small. We observed the same behavior for a significant number of different orders on the sources (figures not presented here), which indicates that these observations do not depend on the order in which the sources are considered.

Figure 2 (right) presents the average degree as a function of the number of destinations considered, for different numbers of sources. After a rather strong initial variation, these plots seem to reach a plateau. Notice that they are not quite constant however: the plots corresponding to a small number of sources seem to decrease slightly as the number of destinations increases, while the plots corresponding to a large number of sources seem to increase slightly.

A similar change is observed for the corresponding densities: for a small number of destinations, considering all sources lead to a smaller density than

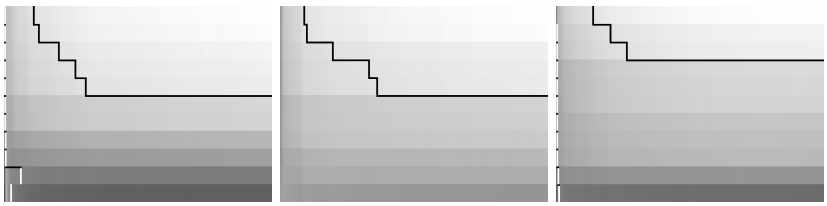

Figure 3: Average degree. Three random orders.

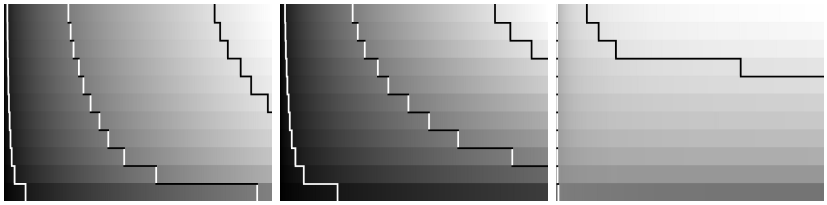

Figure 4: Left: number of nodes. Middle: number of links. Right: average degree. Average grayscale plots for 10 different random orders.

considering a smaller number of sources. Then this trend is inverted as the number of destination grows, and when all destinations are considered, considering more sources leads to a denser graph.

This is probably due to the fact that the observed topology from one source to several destinations is tree-like. When one adds destinations, this adds branches to this tree-like structure, therefore the average degree does not change much, but the density decreases [7]. When the number of destinations is low, sources and destinations play similar roles, and adding sources *or* destinations leads to a decrease of the density. When the number of destinations is high, however, adding sources leads to a densification of the tree-like structure (whereas adding destinations still leads to a decrease in the density).

Figure 3 shows three grayscale plots for the average degree, corresponding to three different (random) orders on sources and destinations. As we can see, they are not identical: the gray is more uniform for the middle plot than for the other two (the 0.5-level line does not appear), which means that the estimation is more accurate: the difference between the maximum and minimum observed average degree is low. The number of sources and destinations needed to achieve a certain precision of the estimation also varies: the left and middle plots reach the 0.9 level line with less sources and destinations than the right plot.

Figure 4 shows the average grayscale plots on 10 different random orders for the number of nodes (left), the number of links (middle) and the average

[7] In a tree, the average degree tends to 2 as the number of nodes grows, whereas the density tends to 0.

degree (right). We first observe that the plot for the average degree is similar to the plots for single orders (Figure 3), which are themselves similar to each other (though not identical). This means that in this case, the obtained value does not strongly depend on the *choice* of sources and destinations: for a given number of sources and destinations, the obtained values are more or less the same for the three plots of Figure 3. In this case, the grayscale plot for the average value over 10 random orders is therefore meaningful, and representative of what one may expect to observe in practice.

The number of nodes, links, and the average degree all increase as the number of sources and destinations grows. This is obvious for the number of nodes and the number of links: more sources and destinations lead to the observation of more nodes and links. The case of the average degree is different, as increasing the number of sources and destinations may increase the number of links less than the number of nodes, causing a decrease in the average degree. Figure 2 (right) shows indeed that, for a small number of sources, the average degree decreases when the number of destinations increases. However, once at least 3 sources are considered, the average degree increases both with the number of sources and destinations.

We can see that the average degree is better estimated than the number of nodes or links. The fact that the average degree is obtained by dividing two other properties (number of nodes and links) which are improved by the use of more sources and destinations has important consequences. If the two properties have the same kind of bias, the quotient may not suffer from this bias: the estimation 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 poorly estimated. This is in accordance with the observations made on models in [14, 15].

4.2. Global and local clustering

The clustering of a graph can be defined in two ways. The first definition is the global clustering, also named transitivity ratio. This is the probability that two nodes are linked, given that they are both connected to a same third node:

$$gc = \frac{3N_{\triangledown}}{N_{\vee}},$$

where $N_{\triangledown}$ denotes the number of triangles (a triangle consists of three nodes connected by 3 links) and $N_{\vee}$ denotes the number of connected triples (i.e. three nodes connected by at least two links) in the graph.

The second definition is the local clustering. The local clustering of a node v (of degree at least 2) is the probability for any two neighbors of v to be linked together:

$$lc(v) = \frac{2.|E_{N(v)}|}{d(v).(d(v)-1)},$$

where $E_{N(v)}$ is the set of links between neighbors of v. Notice that it is the density of the neighborhood of v, and in this sense it captures the local density.

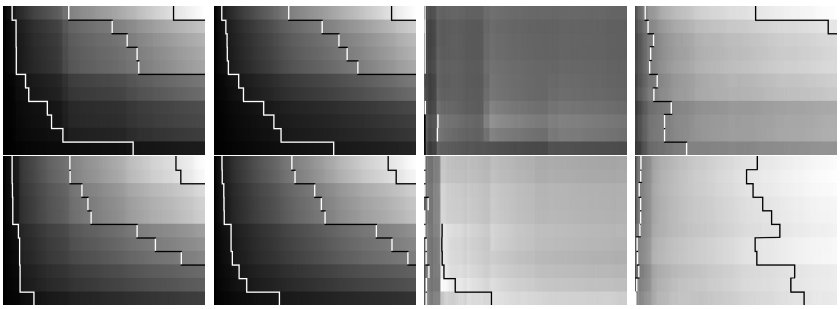

Figure 5: From left to right: number of triangles, number of connected triples, global clustering and local clustering. Each row corresponds to a single random order on sources and destinations.

Then the local clustering of the graph is the average of this value for all nodes (with degree at least 2):

$$lc = \frac{1}{|v \in V, d(v) \geq 2|} \sum_{v, d(v) \geq 2} lc(v).$$

The clustering is strongly related to the numbers of triangles and connected triples in the graph, just as the average degree depends on the numbers of nodes and links (see section 4.1). We will therefore study these properties at the same time.

Figure 5 presents grayscale plots for the number of triangles, the number of connected triples, the global and the local clustering, for two different (random) orders on sources and destinations. We can see that the plots for both clustering notions are different for these two orders, meaning that the choice of sources and destinations strongly impacts them. This is not the case for the number of triangles and connected triples.

We now focus on the difference between the local and global clustering, whose variations are quite different. The local clustering generally increases (the gray becomes lighter) with the number of sources and destinations. This means that it is always under-estimated, and increasing the number of sources and destinations brings it closer to its actual value. The order on sources and destinations plays a strong role on the speed of its evolution: for the second order presented in Figure 5, the 90%-level line is reached with a single source, and the gray does not change much when the number of monitors increases. This means that in this case, the first source in the order gives a good estimation of the local clustering. For the other order, the 90%-level line is only reached with 10 sources (and a large number of destinations).

However the global clustering tends to be over-estimated, whatever the order. The maximum value is reached for small numbers of sources and destinations, and can be quite different from the actual value, see Table 1.

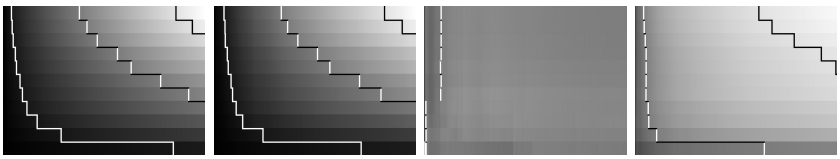

Figure 6: From left to right: number of connected triples, number of triangles, global clustering and local clustering. Average for 10 different random orders.

Figure 6 shows (from left to right) the average grayscale plots for the number of triangles, the number of connected triples, the global and the local clustering. The grayscale plot for the global clustering is quite uniform, which means that the values are very close to each other [8], independently of the number of sources and destinations. Therefore the global clustering seems to be well-estimated (notice that it is slightly over-estimated, as it tends to decrease when the number of sources or destinations increase). Observations made in [14, 15] show that the global clustering is overestimated in graphs with low clustering when the number of sources is low compared to the number of destinations. In this case, explorations discover proportionally more triangles than connected triples. This over-estimation is not a very important one, though, therefore we expect the clustering of a topology measured with more sources and destinations to have a higher global clustering than a random graph. We also noticed that the observed value for the global clustering depends strongly on the choice of sources and destinations. Therefore, the grayscale plot for the global clustering in Figure 6 is not representative of what one may obtain in practice for a single order, and this conclusion must therefore be considered with care.

The average grayscale plot for the local clustering shows that increasing the number of sources and destinations makes its estimation better, as was already the case for single random orders. The average behavior is therefore similar to the behaviors observed for single orders. The wide difference between the plots obtained for different orders however shows that the average plot is only representative of the global behavior of this statistics.

In conclusion, the choice of sources and destinations influences the local and global clustering differently. The local clustering depends more on the number of sources and destinations than the global clustering. Therefore it can be improved by adding sources and destinations, even though the influence of the order on this is very high. The global clustering is more influenced by the choice of the sources and destinations than by their number, meaning that it is very difficult to estimate it accurately.

The authors of [14, 15] considered only the global clustering coefficient, and did not study the impact of the order on sources and destinations. Our observa-

[8] The value of the gray represents the difference with the maximum value observed for the ten considered orders, which is why no white point appears.

Figure 7: Average distance. Left and middle: two different random orders. Right: average of 10 random orders.

tions concerning average grayscale plots for the global clustering are consistent with theirs.

4.3. Average distance

We denote by $d(u, v)$ the distance between two nodes u and v, i.e. the number of links on a shortest path between them. We denote by:

$$d(u) = \frac{1}{n-1} \sum_{v \neq u} d(u, v)$$

the average distance from u to all nodes, and by:

$$d = \frac{1}{n} \sum_{u} d(u)$$

the average distance in the graph.

Notice that these definitions only make sense for connected graphs. In practice, if the graph is not connected (as in the present case) one generally restricts the computation to the largest connected component, which is reasonable since the vast majority of nodes are in this component. The average distance is one of the most classical properties used to describe real-world complex networks and the internet topology in particular. Computing it is however time-costly. To quicken the computations, we use here the heuristics proposed in [17]. It consists in approximating the average distance by choosing at each step i a random node v_i, computing its average distance to all other nodes $d(v_i)$ in time $\mathcal{O}(m)$ and space $\mathcal{O}(n)$, and using it to improve the current approximation. The i-th approximation of the average distance is $d_i = 1/i \sum_{j=1}^{i} d(v_j)$. We stop as soon as the variation in the estimations becomes less than a given ϵ, i.e. $|d_{j+1} - d_j| < \epsilon$. The variable ϵ is a parameter allowing to tune the quality of the approximation vs. the computation time. We use here $\epsilon = 0.1$.

Figure 7 presents grayscale plots for the average distance. We can see that, when one uses few sources, the average distance is over-estimated. The evaluation becomes more accurate when the number of sources and destinations grows. This can be understood as follows: with few sources the graph is close to a tree,

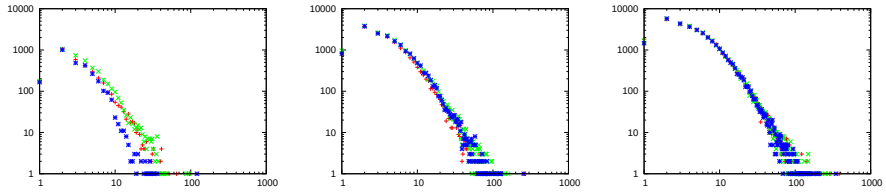

Figure 8: Impact of the choice of sources and destinations on the degree distribution. Each plot presents the degree distributions for three graphs $G_{S'D'}$ obtained with different choices of sources and destinations. Left: S' contains 18% of sources and D' 6% of destinations. Middle: S' contains 45% of sources and D' 26% of destinations. Right: S' contains 81% of sources and D' 67% of destinations.

and the average distance is therefore over-estimated. It changes quickly when one adds more sources.

We observe some fluctuations of the gray level for small numbers of sources and destinations. Once a certain number of sources and destinations is reached, the gray color becomes almost uniform, which means that the average distance does not vary much after this point, and that it is well estimated.

Finally, the impact of the order on sources and destinations on the average distance is very small. There is a strong similarity between the results for different single orders. Figure 7 (left and middle) presents the grayscale plots for two different random orders. We can see that they are very similar, which is also the case for other random orders we considered (not represented here). This means that the choice of sources and destinations has little impact on the observed average distance, and that the average grayscale plot is representative of what one may obtain in practice.

In conclusion we obtain a good estimation of the average distance as soon as a certain number of sources (and destinations) is reached. After this, the average distance becomes accurate and close to the original one. We also showed that the impact of the choice of sources and destinations on the average distance is not important.

This is in accordance with the observations made in [14, 15] on different types of graph models.

4.4. Degree distribution

The degree distribution of a graph is the fraction p_k of nodes of degree exactly k in the graph, for all k [9]. Degree distributions may be homogeneous (all the values are close to the average, like in Poisson and Gaussian distributions), or heterogeneous (there is a huge variability among degrees, with several orders of magnitude between them). When a distribution is heterogeneous, it makes sense to try to measure this heterogeneity rather than the average value. In

[9] Equivalently, one may study the *number* of nodes with degree k.

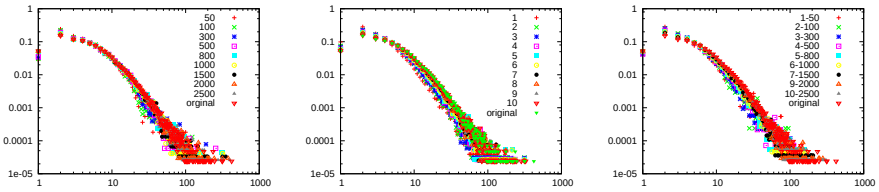

Figure 9: Impact of the number of sources and destinations on the degree distribution. Left: degree distributions of graphs $G_{SD'}$ for which the number of destinations $d = |D'|$ varies from 100 to 3 000 (all 11 sources are considered). Middle: degree distributions of graphs $G_{S'D}$ for which the numbers of sources $s = |S'|$ varies from 1 to 11 (all 3 000 destinations are considered). Right: degree distributions of graphs $G_{S'D'}$ for which both the number of sources and the number of destinations vary.

some cases, this can be done by fitting the distribution by a power-law, i.e. a distribution of the form $p_k \sim k^{-\alpha}$. The exponent α may then be considered as an indicator of how heterogeneous the distribution is. As fitting distributions to power-law leads to disputable results when the distribution is not a perfect power-law, we will however not attempt it here. We will therefore study whether the observed distributions are heterogeneous or not, and whether they are similar to each other.

The degree distribution of the internet is one of the properties for which the bias induced by the exploration has been the most widely studied [1, 2, 6, 7, 8, 9, 10, 16, 17, 21, 23]. Certainly one of the most surprising results is from [16] which shows that power-law distributions can be observed by performing `traceroute` explorations on graphs with an underlying topology following a Poisson degree distribution. The authors of [14, 15] deepen this result by considering several network models and varying the number of sources and destinations. They show that this observation depends of 2 parameters: the underlying topology and the number of sources used to explore it. The exploration of a random graph with a few sources may indeed lead to the observation of a heterogeneous degree distribution. This effect tends to disappear quickly when the number of sources grows. However even a small number of sources in a topology with a power-law degree distribution provides a power-law degree distribution.

We first show to which extent the choice of sources and destinations influences the observed degree distribution. Figure 8 shows the degree distributions for different numbers of sources and destinations. For each case, we present the distributions obtained with three different random choices of sources and destinations. For small numbers of sources and destinations (Figure 8, left), we observe a difference only for degrees larger than ten, whereas for low degrees the distributions are almost identical. When the number of sources and destinations increases (Figure 8, middle and right), these distributions tend to become very similar for all values of the degree. This shows that the choice of sources and destinations does not influence much the observed degree distribution, provided their number is not too small.

Since the choice of sources and destinations does not influence much the obtained degree distribution, we can now study the impact of the *number* of sources and destinations, without worrying about their choice. We increase the number of sources and destinations separately to show their impact on the degree distribution. Figure 9 shows the degree distributions for graphs in which we vary the number of destinations (left), the number of sources (middle), and both at the same time (right). The first observation is that these degree distribution are all heterogeneous and do not vary greatly. This confirms observations made by simulations in [14, 15].

The distributions of Figure 9 (left) coincide more precisely than the others. This indicates that the degree distribution is more accurately estimated when the number of sources is high. In this case, changing the number of destinations does not alter the distribution much.

The distributions of Figure 9 (middle and right) converge to the degree distribution of the final graph as the number of sources and destinations increases. Since we saw that the number of destinations does not play a great role in the degree distributions, this means that the change is driven by the number of sources.

In summary, we have a good approximation of the degree distributions, even with a small number of sources and destinations. It becomes even more accurate as the number of sources grows. It is important to notice that in all cases, the observed degree distribution is heterogeneous. This confirms observations made by simulations [14, 15].

5. Related work

Since a few years, many works have conducted experimental and formal studies to evaluate the accuracy of the obtained maps of the internet. Most of these studies focus on the impact of the measurement procedure on the obtained degree distribution, and use models for the topology and the `traceroute` exploration to evaluate this [1, 2, 6, 7, 8, 9, 10, 14, 15, 21, 23]. They give good arguments for the fact that the maps of the internet collected with a single source are very incomplete, and probably suffer from an important bias. Many of these works agree that increasing the number of sources quickly reduces this bias, at least concerning the degree distribution.

Some works have addressed specifically the question of the impact of the number of sources and destinations on the obtained data: [5] and [26] study the number of nodes and links discovered as a function of the number of sources and destinations used for the exploration. They both exhibit a *diminishing returns* effect, each new source adding less information than the previous one. As explained in Section 3, both papers added sources approximately in the order given by the greedy maximum heuristics. Studying also random orders allowed us to moderate their observations. [14, 15] study the impact of the degrees of sources and destinations on the observed number of nodes and links.

The authors of [26] study the impact of the number of sources on a number of graph properties, using data collected from the distributed measurement project Dimes [24]. They consider sources by decreasing order of the number of links they discover, and do not consider the impact of the *choice* of sources on these properties.

The grayscale plots were introduced in [14, 15] in order to obtain an in-depth understanding of the impact of the number of sources and destinations on a number of widely studied graph properties. They used models for different types of networks in order to study the impact of the network topology on the observations. [5] also used contour plots, which are somewhat similar to grayscale plots, for studying the impact of the number of sources and destinations on the observed number of nodes and links. Our contribution is complementary to these works in two ways: first, we use real data, whereas these works study models both for the networks and the `traceroute` exploration ([5] uses real data but only studies the observed number of nodes and links); second, in this paper we study not only the impact of the *number* of sources and destinations, but also their *choice*.

Finally, some works propose empirical criteria for identifying whether the observed properties can be trusted [17, 14, 15, 26, 16]. This is similar to our approach.

6. Conclusion

We conducted an extensive set of experiments aimed at evaluating the impact of the sources and destinations used in `traceroute`-like measurements on the properties of the observed topology. Our goal was to estimate whether the observed properties are the actual properties of the topology, or if they are biased by measurement artifacts.

We used real data obtained from `traceroute`-like measurements, so our results do not rely on simulations. This has the advantage that we did not need to rely on models, either for the internet topology, of for the `traceroute` measurements.

As expected from previous work in this area, we showed that the number of sources and destinations plays a strong role on the observed properties. We also studied in depth the impact of the *choice* of sources and destinations on the observed properties, and showed that it is important. When one uses two different sets of sources and destinations with the same size, one may obtain graphs which are very different from each other. This is true even for basic properties such as the number of nodes and links.

We studied various graph statistics which are widely used for graph description. We showed that they do not behave in the same way with respect to changes in the sources and destinations used for the exploration. Some of them are strongly dependent on the choice of sources and destinations and/or their number. Such properties can therefore not be trusted, since performing measurements with different sets of sources and destinations would lead to very

different results. This is the case for instance for the clustering coefficient. On the other hand, some properties are very resilient to changes in the sets of sources and destinations, and are therefore probably accurate descriptions of the actual internet topology. This is the case for instance for the average distance and the degree distribution.

This works could be extended in several directions. We showed that one may trust in some properties more than in others: the average distance is probably accurately evaluated, while the clustering coefficient most probably is not, for instance. This should be compared to observations on other data sets. We also showed that some properties which are the ratio of two other properties may sometimes be well estimated, even if the two base properties are not accurate: if they suffer from the same bias, this bias is removed by the ratio. It would be beneficial to detect other properties that can be accurately estimated. Such properties would probably not be ones that are classically used, since we studied most of these in this paper. However the fact that they can be accurately estimated, while more classical properties cannot be trusted, would increase their interest. In particular, the authors of [17] noted that the ratio between the clustering coefficient and the density was probably more accurate than any of these properties. This idea should be explored further.

We observed that different sets of sources and destinations lead to statistically different views of the topology. Designing methods for deciding which sources and destinations will provide a representative view of the topology before the measurements start is therefore an interesting goal. This would however still probably require some preliminary experimental measurements, as some prior knowledge seems necessary in order to know how different sources will complement each other. Since the choice of sources is limited by the fact that one must have access to the corresponding computers, a relevant approach might be to assign different destination sets to different sources, in order to increase the contribution of each source to the global representativeness.

Finally, the internet topology is not a static object, and it evolves with time, see for instance [19]. Performing several rounds of measurements, as was done for the data we use here, therefore aggregates up-to-date data with obsolete data. This has certainly an impact on the properties of the obtained topology (for instance, one may expect that it increases the density of the obtained graphs). Taking this dynamics into account in the evaluations of the properties of the observed topology is therefore an important question.

References

[1] D. Achlioptas, A. Clauset, D. Kempe, and C. Moore. On the bias of traceroute sampling; or, power-law degree distributions in regular graphs. In *ACM Symposium on Theory of Computing (STOC2005)*, 2005.

[2] D. Achlioptas, A. Clauset, D. Kempe, and C. Moore. On the bias of traceroute sampling. *Journal of the ACM*, 56(4):1–28, 2009.

[3] Caida – archipelago project. `http://www.caida.org/projects/ark/`.

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

[5] Paul Barford, Azer Bestavros, John Byers, and Mark Crovella. On the marginal utility of network topology measurements. In *Proceedings of the 1st ACM SIGCOMM Workshop on Internet Measurement, in conjunction with Internet Measurement Conference*, 2001.

[6] Qian Chen, Hyunseok Chang, Ramesh Govindan, Sugih Jamin, Scott Shenker, and Walter Willinger. The Origin of Power-Laws in Internet Topologies Revisited. In *IEEE Infocom*. IEEE, 2002.

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

[8] R. Cohen, M. Gonen, and A. Wool. Bounding the bias of tree-like sampling in IP topologies. *Networks and Heterogeneous Media*, 3:323 – 332, 2008.

[9] R. Cohen and D. Raz. The Internet Dark Matter - on the Missing Links in the AS Connectivity Map. In *Proceedings of IEEE INFOCOM 2006. 25TH IEEE International Conference on Computer Communications*, pages 1–12, April 2006.

[10] L. Dall'Asta, J.I. Alvarez-Hamelin, A. Barrat, A. Vázquez, and A. Vespignani. A statistical approach to the traceroute-like exploration of networks: theory and simulations. In *Workshop on combinatorial and Algorithmic Aspects of Networking (CAAN)*, 2004.

[11] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationships of the Internet topology. In *SIGCOMM*. ACM, 1999.

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

[13] J.-L. Guillaume and M. Latapy. Complex networks metrology. In *Complex systems*, 2005.

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

[15] Jean-Loup Guillaume, Matthieu Latapy, and Damien Magoni. Relevance of massively distributed explorations of the internet topology: Qualitative results. *Computer Networks*, 50 (16):3197–3224, 2006.

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

[17] M. Latapy and C. Magnien. Complex network measurements: Estimating the relevance of observed properties. In *IEEE infocom*, 2008.

[18] M. Latapy, C. Magnien, and F. Ouédraogo. A radar for the internet. In *Proc. first International Workshop on Analysis of Dynamic Networks (ADN), in conjunction with IEEE ICDM 2008*, 2008. Available at `http://arxiv.org/abs/0807.1603`.

[19] Clémence Magnien, Frédéric Ouédraogo, Guillaume Valadon, and Matthieu Latapy. Fast Dynamics in Internet Topology: Observations and First Explanations. In *2009 Fourth International Conference on Internet Monitoring and Protection*, pages 137–142. IEEE, 2009.

[20] Damien Magoni and Mickaël Hoerdt. Internet core topology mapping and analysis. *Computer Communications*, 28(5):494–506, 2005.

[21] T. Petermann and P. De Los Rios. Exploration of scale-free networks. *Eur. Phys. J. B*, 38(2), 2004.

[22] Radar program and data. `http://www-rp.lip6.fr/~latapy/Radar`.

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

[24] Y. Shavitt and E. Shir. DIMES: Let the internet measure itself. *ACM SIGCOMM Computer Communication Review*, 35(5), 2005. See `http://www.netdimes.org`.

[25] Caida – skitter project. `http://www.caida.org/tools/measurement/skitter/`.

[26] Udi Weinsberg and Yuval Shavitt. Quantifying the importance of vantage points distribution in internet topology measurements. In *Proceedings of IEEE Infocom*, 2009.