Detecting Events in the Dynamics of Ego-centered Measurements of the Internet Topology [1]

Matthieu Latapy, Assia Hamzaoui, Clémence Magnien

LIP6 – CNRS and UPMC

Firstname.Lastname@lip6.fr

Abstract

Detecting events such as major routing changes or congestions in the dynamics of the internet topology is an important but challenging task. We explore here an empirical approach based on a notion of statistically significant events. It consists in identifying properties of graph dynamics which exhibit a homogeneous distribution with outliers, corresponding to events. We apply this approach to ego-centered measurements of the internet topology (views obtained from a single monitor) and show that it succeeds in detecting meaningful events. Finally, we give some hints for the interpretation of detected events in terms of network operations.

Keywords: internet, measurement, dynamics, graphs, event, monitoring

1 Introduction

The internet nowadays plays a key role in our society, economy, and everyday life. Despite this, our knowledge of the problems experienced by its infrastructure (failures and attacks, congestions, bugs, etc) remains very limited. As a consequence, when we face major losses of connectivity or service degradations we have a limited understanding of the underlying phenomena, their impact on the internet, and the solutions to prevent them.

Indeed, in this context, monitoring the network at a global scale remains out of reach. The first challenge to be addressed would be to obtain a global view of the network, which is generally done using `traceroute`-like tools and/or BGP tables. Such maps are however very partial and biased, and their collection is too expensive (in time and network load) to be repeated at a high-enough frequency for monitoring the network in real-time.

To solve these issues, it was proposed to focus on a part of the topology called an *ego-centered view*. It consists in what a single machine, called *monitor*, may see of the internet topology. It is basically captured by running `traceroute` measurements from the monitor to a given set of randomly chosen targets, and iterating this process every few minutes. This may be done very efficiently, and measurements of this kind have been performed at a large scale. See [27] for details. The main outcomes of these works are the obtained dataset [39] and the observation that the dynamics of ego-centered views is much higher than expected [30].

It must be clear that such measurements give a very partial view of the global topology. In addition, the ego-centered nature of these views has a strong impact on observations. In particular, in case of a connectivity loss at the monitor (or close to it), the view becomes blank

[1] An abstract of this work was presented at the *Workshop on Dynamic Networks*, WDN 2010 (same title, same authors).

(or almost), which may lead to the conclusion that a major problem occurred on the internet. This is not true, though, as the connectivity loss may be of very limited scope (possibly only the monitor). Still, such measurements have the avantage that they have a natural and easy to understand meaning (they are basically routing trees from the monitor), and that they can be repeated at a relatively high frequency (of the order of a few minutes typically).

Even if precise data were available regarding the dynamics of the internet topology (perfectly accurate maps captured every second, for instance), monitoring this dynamics would still face another difficult challenge: what are *events* in the dynamics of such objects, and how to detect them? Given our very limited knowledge of the dynamics of the internet, and because this dynamics is intense (due in part to load-balancing [3, 32, 37], BGP instability [24, 25] and other phenomena [30]), answering such questions raises fundamental issues.

In addition, the object under concern is a *graph*, *i.e.* a set of nodes and links beween them, which evolves during time. Whereas many works consider temporal series of *numbers* and detection of events within them, only very few previous works consider dynamic *graphs*, and none addresses the definition and detection of events in their dynamics. Current state-of-the-art on the analysis of such dynamics indeed remains very limited. In order to detect events in the dynamics of the internet topology, one therefore has to introduce graph-based notions of events, together with a method to detect them.

We propose in this paper a general empirical method for event detection in graph dynamics and demonstrate its effectiveness by applying it to ego-centered measurements of the internet topology. We start with such measurements [39] (Section 2) and we consider a notion of *statistically significant events* with a method to detect them (Section 3). We then introduce simple and more subtle dynamic graph properties and assess their relevance for event detection in our data (Sections 4 to 6). We finally discuss methods for interpretation of detected events (Section 8) and correlations between detected events (Section 8.1). We compare our approach to previous works in Section 9 and conclude in Section 10.

2 Data and notations

We present here the data we used in the whole paper and mathematical notations for manipulating them.

2.1 Data

Although it is sometimes inaccurate [3], the `traceroute` tool [21] basically gives an IP-level path from the monitor running it to the selected target. Each node on this path is an IP address and each link represents an IP-level hop. This tool is at the basis of most IP-level internet topology measurements, and maps are constructed by merging such measured routes from several monitors to many targets, see for instance [41, 2, 20, 15, 44].

The `tracetree` tool [27] works in a very similar way but gives a routing tree from a monitor to a set of targets, which is called an ego-centered measurement because it gives a view of the internet from this specific monitor. It is basically equivalent to running `traceroute` from the monitor to each target in the set. Compared to this approach, it however imposes lower and more balanced load on the network, and is faster. As a consequence, it may be iterated at a relatively high frequency, leading to *radar* measurements. Such measurements consist in series of periodic ego-centered measurements, each being called a *round*.

Radar measurements are presented in [27] and the authors provide the obtained data freely [39]. They ran these measurements from more than a hundred nodes towards 3,000 random targets each, during several weeks in continuous and with approximately one round every 15 minutes. Each monitor therefore provides its own ego-centered view of the dynamics of the internet IP-level topology. The method we present here considers such a view, obtained from a given monitor, and detects events in the dynamics it captures. We performed our computations on views from a wide set of monitors and obtained similar observations. As our goal here is not to discuss subtle differences between monitor views, we present the results for a representative monitor only, located in Japan (Tokyo University), which performed 5000 rounds of measurement (during more than 7 weeks in spring 2007). Comparing observations from several monitors more precisely is one of our main perspectives, see Section 10.

Finally, notice that, like the ones of `traceroute`, the probes sent by `tracetree` do not necessarily receive an answer. This leads to unidentified nodes on the paths (traditionally represented by stars '*'). Matching unidentified nodes from one round to another is a difficult problem, and may interfere with event detection. We therefore decided to simply ignore them (*i.e.* remove them from the measurement) and to focus on nodes with an actual IP address. In addition, we removed isolated nodes (nodes with no link), which provide no topological information. This is classical in analysis of such internet measurements.

2.2 Notations

We consider series of rounds of measurements performed from a monitor M, indexed by an integer i. Round i corresponds to a tree $G_i = (V_i, E_i)$ which we consider as an undirected graph: V_i is a set of nodes (IP addresses) and $E_i \subseteq V_i \times V_i$ is a set of links. We denote by $N_i = |V_i|$ the number of nodes at round i.

For any two integers i and j, we denote by $G_i^j = (V_i^j, E_i^j)$ the graph obtained by merging rounds from i to $i + j - 1$ (*i.e.* j rounds starting from the i-th): $V_i^j = \cup_{k=i}^{k=i+j-1} V_k$ and $E_i^j = \cup_{k=i}^{k=i+j-1} E_k$.

For any integer i, given two integers p and c we call $G_i^c = (V_i^c, E_i^c)$ the *current* graph and $G_{i-p}^p = (V_i^{i-p}, E_i^{i-p})$ the *previous* graph. The current graph is the merging of the c rounds starting from the i-th, and the previous graph is the merging of the p rounds preceding the i-th.

We will use these notations in next sections to define properties of dynamic graphs.

3 Methodology

As explained above, only very limited knowledge is available regarding the dynamics of the internet topology, and of any dynamic graph in general. Therefore, one cannot rely on a description of the *normal* dynamics to define events as behaviors which deviate from the normal one. Likewise, there is no clear knowledge of event occurrences and their impact of the dynamics of ego-centered views. Therefore one cannot rely on the study of such known events and search for their signature in observed dynamics. For these reasons, we propose a statistical approach to event detection. It relies on the definition of dynamic graph properties (Sections 4 to 6) and identification of statistically significant deviations in the evolution of these properties. We detail this approach here.

First notice that when one considers a set of numerical value associated to a dynamic property (like for instance the number of nodes in a dynamic graph), three typical situations may occur: observed values may be homogeneous, which means that they are all similar to the average value and that one never observes any significant deviation from it; observed values may be heterogeneous by nature, which means that there is no notion of a *normal* value; and observed values may be homogeneous with some outliers, *i.e.* most values have a homogeneous nature but some significantly deviate from them.

In the first two situations, the considered property is in itself of no help for event detection: either all values are normal and there is no notion of event (homogeneous values), or there is no notion of normal behavior and thus no event (heterogeneous values). In the third case, on the contrary, the property may be used for event detection: statistically significant outliers indicate events, while most values correspond to the normal behavior.

We use this notion here. Therefore, we are interested in properties with homogeneous values and outliers. This leads to the following four step method:

1. define numerical properties describing the dynamics of the object under concern,
2. compute them on the considered dataset,
3. study the distribution of obtained values (*i.e.* number of occurrences of each possible value) and decide on the nature of the property: homogeneous, heterogeneous or homogeneous with outliers,
4. select the properties which are homogeneous with outliers and consider that these outliers point out events in the dynamics.

It must be clear that defining relevant properties which capture the dynamics of the object and have the expected behavior (homogeneous values with outliers) is a challenge in itself; we address it in the following sections. Another difficult task is the study of empirical distributions and the decision regarding their nature. In principle, this should be feasible using classical fit methods and goodness of fit evaluations. In practice, though, empirical distributions rarely fit models very well, automatic methods may be misleading [45, 10], and appropriately fitting empirical data is a challenginng task [11]. We therefore combine visual inspection of distributions and automatic statistical techniques, both being complementary of each other and having their own limitations and strengths.

In order to perform visual inspection of distributions, we plot them in lin-lin, lin-log and log-log scales. This makes it possible to observe bell-shaped distributions (visible in lin-lin scales), exponential decreases (straight lines in lin-log scales) and polynomial decreases (straight lines in log-log scales). Exponential decreases in distributions are hallmarks of homogeneity (values are exponentially rarer when they increase) and polynomial decreases are hallmarks of heterogeneity (values are *only* polynomially rarer when they increase). Notice that we assume in this presentation that distributions are decreasing, but similar reasoning is applicable for their increasing parts.

We also study in similar ways the inverse cumulative distributions (*i.e.* for each possible value, the number of occurrences larger than it), which are often easier to read. Examining lin-lin, lin-log and log-log plots of these leads to conclusions in a similar way to what we described above.

Figures 1 to 3 illustrate the different typical situations and their visual inspection. We consider three distributions (one figure each) and plot them in both lin-lin, lin-log and log-log scales (first row, from left to right). We also plot the inverse cumulative distribution in each scale (second row, from left to right), leading to six plots for each distribution. Figure 1 illustrates the homogeneous case: there exists a *normal* value, and no observed value deviates much from it. This is captured by the exponential slope of the distribution, revealed by its straight shape in lin-log scales. Figure 2 illustrates the heterogeneous case: there exists no notion of a *normal value*; instead, the distribution is characterized by the fact that observed values cover wide ranges, with a polynomial decrease revealed by its straight shape in log-log scales. Finally, Figure 3 illustrates the homogeneous case with outliers: there exists a *normal* value, but some values deviate much from it. These values are called *outliers* and indicate statistically significant events. The corresponding distributions therefore exhibit two regimes: an exponential decrease (revealed by a straight shape in lin-log scale) and some values which deviate significantly.

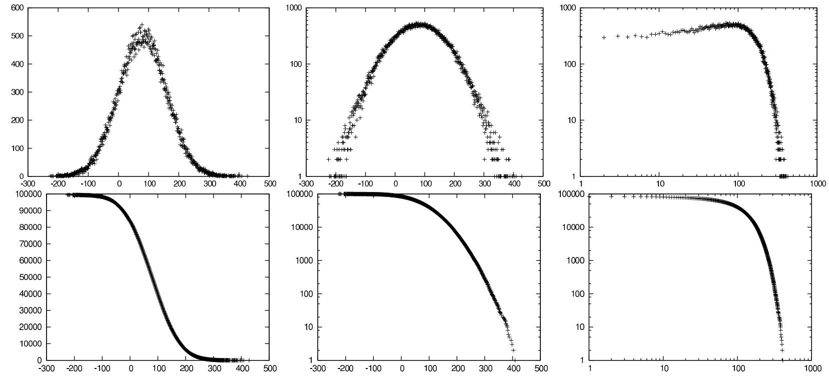


Figure 1: Typical homogeneous distribution. First row (left to right): the distribution in lin-lin, lin-log and log-log scales. Second row: the inverse cumulative of the distribution in the same scales.

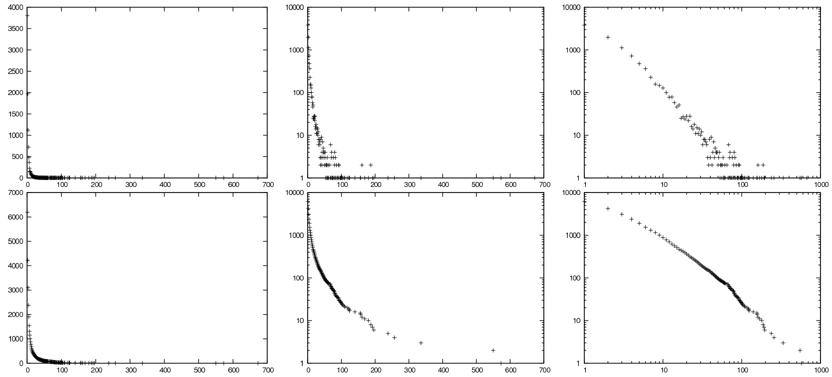


Figure 2: Typical heterogeneous distribution. First row (left to right): the distribution in lin-lin, lin-log and log-log scales. Second row: the inverse cumulative of the distribution in the same scales.

In addition to visual inspection of plots, we use automatic techniques to perform decisions on the types of observed distributions. Such techniques rely on the use of model distributions to which empirical distributions may be fitted. There is a wide variety of possible model distri-

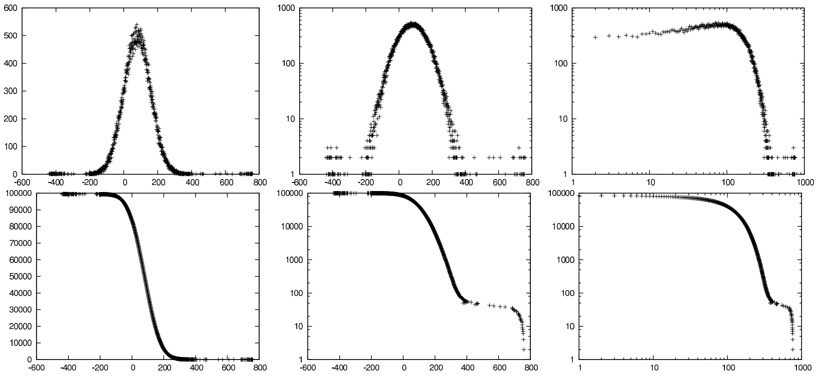


Figure 3: Typical homogeneous distribution with outliers. First row (left to right): the distribution in lin-lin, lin-log and log-log scales. Second row: the inverse cumulative of the distribution in the same scales.

butions, but our goal here is not to find a perfect fit: we want to decide whether the empirical distribution is homogeneous, heterogeneous or homogeneous with outliers, as described above. We therefore consider only two model distributions [2] to capture the homogeneous and heterogeneous nature of an empirical distribution: the normal distribution $P(x) = \frac{1}{\nu\sqrt{2\pi}} e^{\frac{-1}{2}(\frac{x-\mu}{\nu})^2}$, which is a typical homogeneous distribution with well defined mean μ and standard deviation ν and an exponential decrease; and the power-law distribution $P(x) \sim x^{-\alpha}$, which is a typical heterogeneous distribution characterized by its exponent α and which has no *normal* value.

In this framework, deciding on the homogeneous or heterogeneous nature of a given empirical distribution consists in: (1) trying to fit it with each model (*i.e.* determine the model parameters which make it fit the empirical distribution as well as possible); (2) deciding which model corresponds best to the empirical distribution: this the one that produces the fit closest to this distribution. We detail these two steps below.

In order to handle the crucial case where the empirical distribution is homogeneous with outliers, we add an additional step: first we identify and remove possible outliers, and then we compare the obtained distribution to the normal model, in the same way as we compare the original distribution to the normal and power-law models, which we detail below. Identifying possible outliers in a distribution is a difficult problem in itself [43, 18]; we will use here a simple method based on Grubb's test. It consists in comparing each value to the mean and standard deviation of the distribution: if the value is higher or lower than the mean by a given number of times the standard deviation, then it is considered an outlier.

The most classical methods for performing fits probably are to minimize the error (ME) at each point of the empirical distribution and to maximize the likelihood (ML) of the empirical distribution for the model. Here we use the classical Maximum Likelihood Estimation (MLE) [16, 12] because ME uses only the moments of the empirical distribution rather than all the data and is very sensitive to initialization values; since we have no insight for such initialization, this makes it impossible to use for performing automatic decisions.

Notice that, in our context, the fit is not the outcome of greatest interest: we are interested in how much each fit is relevant, in order to compare the fits produced by the different models. In a way consistent to the way the fit is computed (MLE), one could use the obtained likelihood

[2] We conducted the same computations with Poisson and gamma models, leading to very similar results.

to estimate this relevance. However, the notion of likelihood depends on the considered model, and thus comparing likelihoods obtained with different models makes little sense, if any.

We therefore compare the relevance of the fits by comparing the empirical distribution to each fit directly. A classical way to do so is the Kolmogorov-Smirnov (KS) distance [38], which is the largest difference between the cumulative versions of the empirical distribution and the fit. This gives a *worst case* comparison, as the point at which the fit is the poorest entirely determines the value of the KS distance. If the fit is poor at this point but excellent everywhere else, the KS distance will be high. In order to obtain more insight, we also compute the Monge-Kantorovich (MK) distance [17], defined as the average distance between the cumulative versions of the empirical distribution and the fit.

We finally obtain a complete method for automatically deciding whether an empirical distribution has an homogeneous, heterogeneous, or homogeneous with outliers nature: we fit it to normal and power-law model distributions using MLE; we remove outliers with Grubb's test and fit the obtained distribution to the normal model distribution; then we compute the KS and MK distances and conclude that the distribution is homogeneous if it is closest to the fit with the normal model, heterogeneous if it is closest to the fit with the power-law model, or homogeneous with outliers if the distribution with the outliers removed is closest to the normal model.

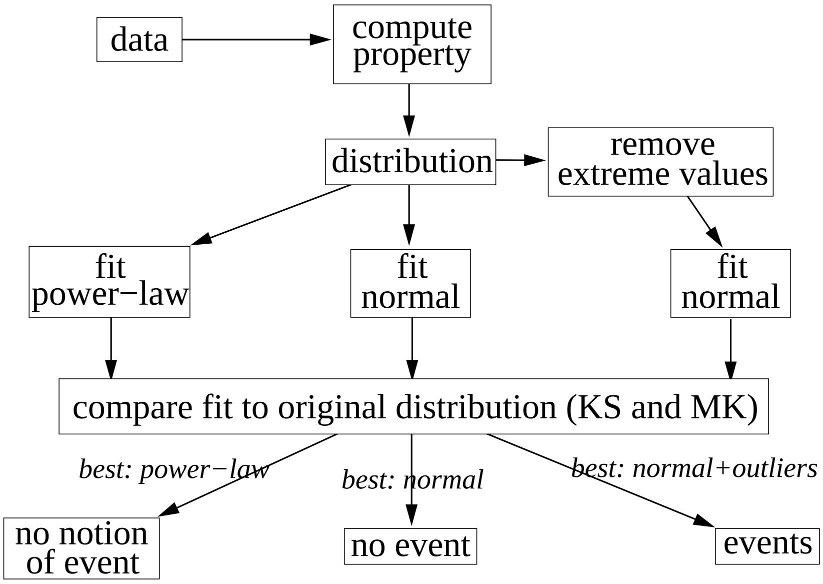


Figure 4: Overview of our method: for each property of possible interest (typically the ones defined in Sections 4 to 6), we compute its values on the dataset, obtain the distribution and fit it to the three model distributions discussed in the text (power-law, Poisson, and Poisson with outilers); depending on the best fit result (according to KS and MK distances), we decide wether the property is able to detect statistically significant events in the dataset or not. Once this method is applied to several properties, we investigate correlations between results (Section 7) and we interpret detected events (Section 8).

Our global method is summarized by the diagram presented in Figure 4.

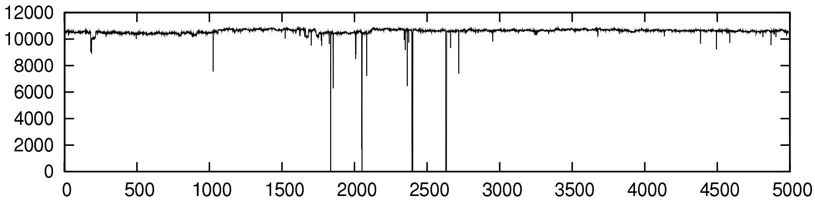
4 Numbers of nodes

The most basic property of ego-centered views certainly is the number of nodes observed at each round of measurement, which we study here. We also consider the number of nodes in several consecutive rounds and the number of appearing nodes (nodes observed in some rounds of measurements but not in previous rounds).

Although these properties are barely graph properties (they do not capture any information on the *structure* of the network), we consider them as a basis for further investigation, as they are the most basic yet nontrivial statistics. We investigate more subtle graph properties in next sections.

4.1 Number of nodes in each round

We plot in Figure 5 the number N_i of nodes observed in round i, as a function of i. This plot shows that the number of nodes at each round is very stable with some exceptions. Most of the time, it oscillates close to a mean value slightly above 10 600. However, closer examination shows that this value changes near rounds 1 100 and 2 100: during some time after these rounds, the number of nodes oscillates close to a different value. In addition to these changes in the average value, the plot also exhibits sharp downward peaks. On the contrary, no upper peak is visible.



	KS	MK
normal	0.3138	0.2394
with outliers	0.1003	0.0993
power-law	0.6215	0.5798

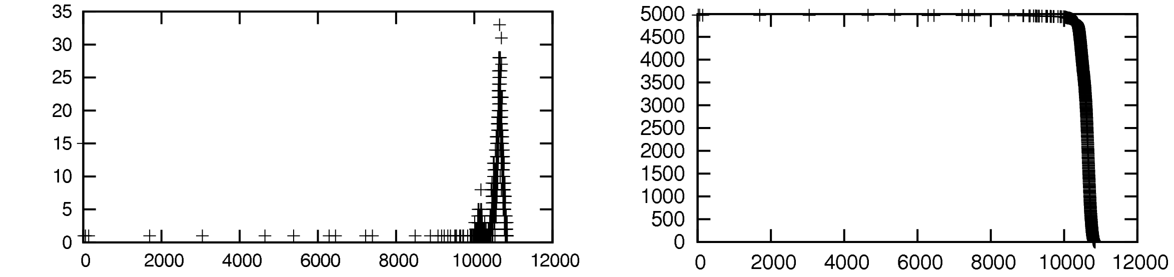
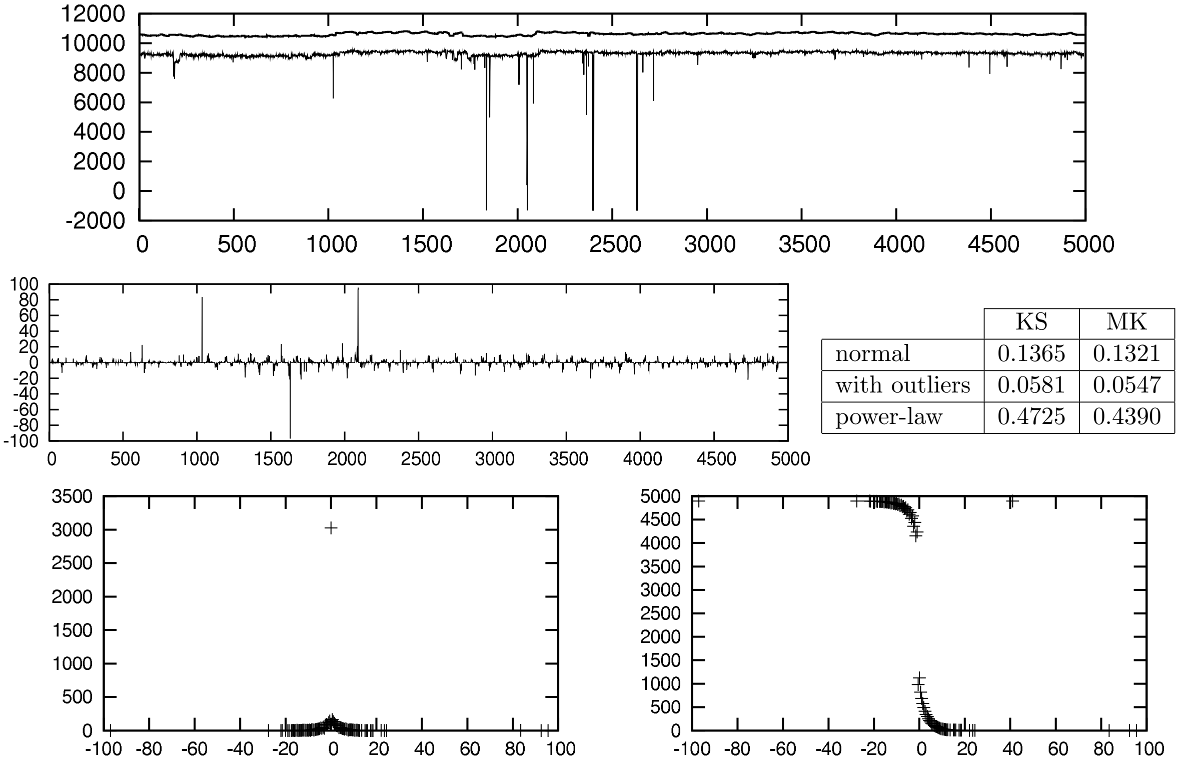


Figure 5: Number N_i of nodes observed at measurement round i, as a function of i, and the distribution and inverse cumulative distribution of this value. The table displays KS and MK tests for the distribution with each considered model distribution.

These observations are confirmed by the distribution of the value of N_i and the goodness of fit test, see Figure 5. Indeed, the distribution reveals two distinct regimes, with many values around 10 050 and 10 600. Otherwise, the distribution is clearly homogeneous with outliers. The presence of abnormally low values (points on the left) but no abnormally high values corresponds the presence of downward peaks but no upward ones. There is also a more unstable regime between 1600 and 1800.

It must be clear that downward peaks, although they are very clear statistical outliers, bring little information: they may be caused by local connectivity failures, which have the effect that ego-centered views are (partly) blank during one or a few rounds. This kind of event is trivial.

On the contrary, an upward peak would indicate an interesting event: it would mean that we suddenly observe significantly more nodes at one round. However, there is no upper peak on this plot, which is a non-trivial fact: one may easily imagine scenario where such peaks would appear. For instance, one may in theory switch from a situation where some paths have a long common prefix to a situation where they have much less nodes in common, thus leading to a significant increase in the number of distinct observed nodes. Figure 5 shows that such scenario do not occur in practice. As a consequence, one cannot detect events by observing abnormally high values of N_i.



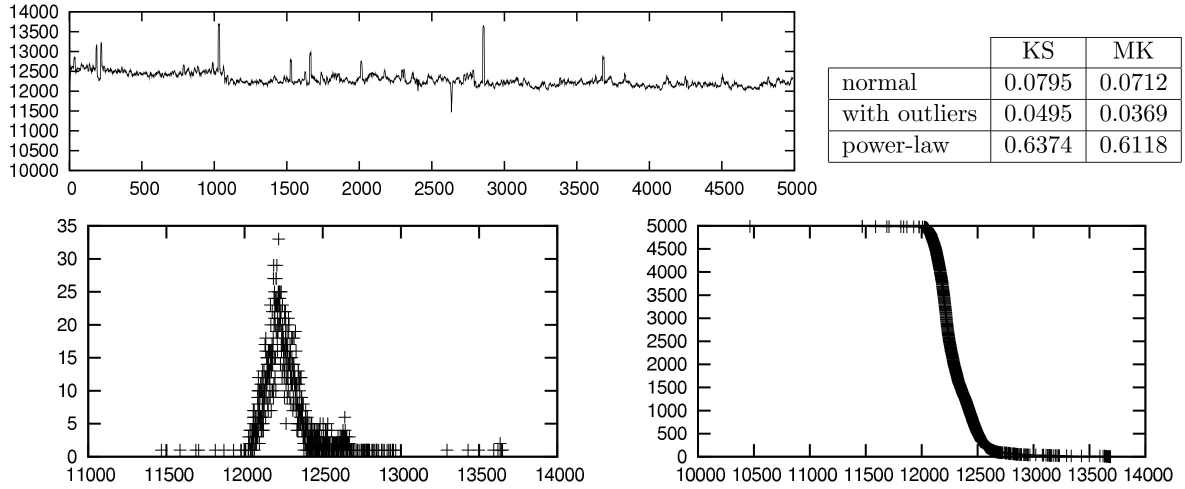
	KS	MK
normal	0.1365	0.1321
with outliers	0.0581	0.0547
power-law	0.4725	0.4390

Figure 6: Top plot: number N_i of nodes observed at measurement round i and its median M_i for 100 values after i, as a function of i (we shifted down N_i for readability). Middle plot: the variations of M_i. Bottom plots: the distribution and the inverse cumulative distribution of these variations. The table displays KS and MK tests for the distribution with each considered model distribution.

Finally, the most notable dynamics in the number N_i of nodes observed at each round are changes in the mean values around which it oscillates. We detect such changes as follows: we associate to each i the median of values N_i to N_{i+100}, that we denote by M_i, then we define consider the variations of the median, *i.e.* $M_i - M_{i-1}$ for all i. We plot these values in Figure 6. It is clear that this statistics succeed (both visually and with our automatic method) in identifying events, *i.e.* outliers in the distribution. This is our first way to detect statistically significant events.

4.2 Number of nodes in consecutive rounds and appearing nodes

The fact that the number N_i of nodes observed at each round is very stable does *not* mean that the observed nodes are always the same: consecutive rounds may see different ones. Such changes may be evidenced by observing the number N_i^c of distinct nodes in c consecutive rounds, for a given integer c. We display in Figure 7 the case $c = 5$.



	KS	MK
normal	0.0795	0.0712
with outliers	0.0495	0.0369
power-law	0.6374	0.6118

Figure 7: Number N_i^5 of distinct nodes observed during at least one of the five rounds preceding the i-th, as a function of i, and the distribution and inverse cumulative distribution of this value. The table displays KS and MK tests for the distribution with each considered model distribution.

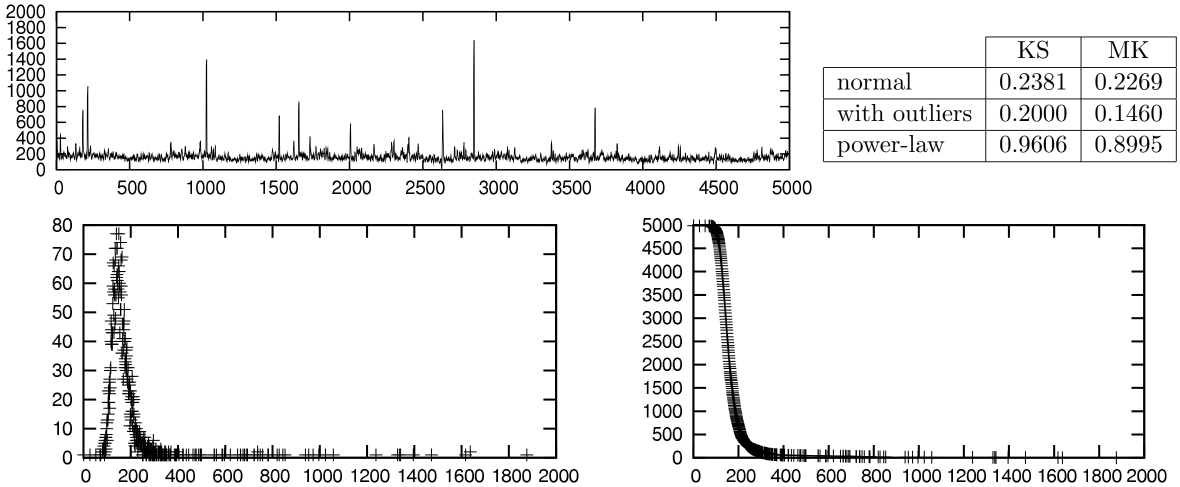
This plot shows that, like N_i, N_i^5 is rather stable and oscillates around a mean value[3]. As expected, this value is larger than the one for N_i, but it is far from 5 times larger. This shows that many nodes appear in several consecutive rounds. Moreover, upper peaks appear on this plot, which make it very different from the one of N_i in Figure 5. The distribution and fit tests it confirm the presence of a clear mean value, but also points out clear statistical outliers, both abnormally low (as before) and abnormally high (which is new).

This observation is important for event detection: there are specific times (pointed out by the peaks in Figure 7) at which an abnormal number of new nodes appear in a series of consecutive rounds. This gives a new way to detect statistically significant events.

However, automatic detection using this approach is not trivial: as the observed mean value may change during time, and as upper peaks (which we want to detect) may be smaller than these variations of the mean, we may miss some events, and making the difference between a statistically sound event and normal dynamics may be difficult. In order to solve this problem, we now define *appearing nodes* as the nodes observed at some point in a series of c consecutive rounds but not observed in any of the p previous rounds, for some integers c and p. With our notations, the appearing nodes at round i are the nodes in $V_i^c \setminus V_{i-p}^p$.

[3] It also experiences changes of regime, like N_i, for instance around round 1100 in Figure 7. Notice that, in this case, the new average value for N_i^5 is larger than before, while it was lower for N_i. This means that, although we see less nodes in each round, the nodes we see vary more from one round to another. This gives some hints on further understanding the event which occurred, but deepening this is out of the scope of this paper.

Notice that observing the number of disappearing nodes is also natural, as well as appearing and disappearing links. We observed similar results for all these notions, and so we focus here on appearing nodes. We also considered wide ranges of possible values for c and p, and observed little difference, if any, as long as they were greater than 1 or 2 and lower than 100. We illustrate here the obtained results with $p = 10$ and $c = 2$, see Figure 8.



	KS	MK
normal	0.2381	0.2269
with outliers	0.2000	0.1460
power-law	0.9606	0.8995

Figure 8: Number a_i of appearing nodes at round i as a function of i, and the distribution and inverse cumulative distribution of this value. The table displays KS and MK tests for the distribution with each considered model distribution. We used representative values $p = 10$ and $c = 2$.

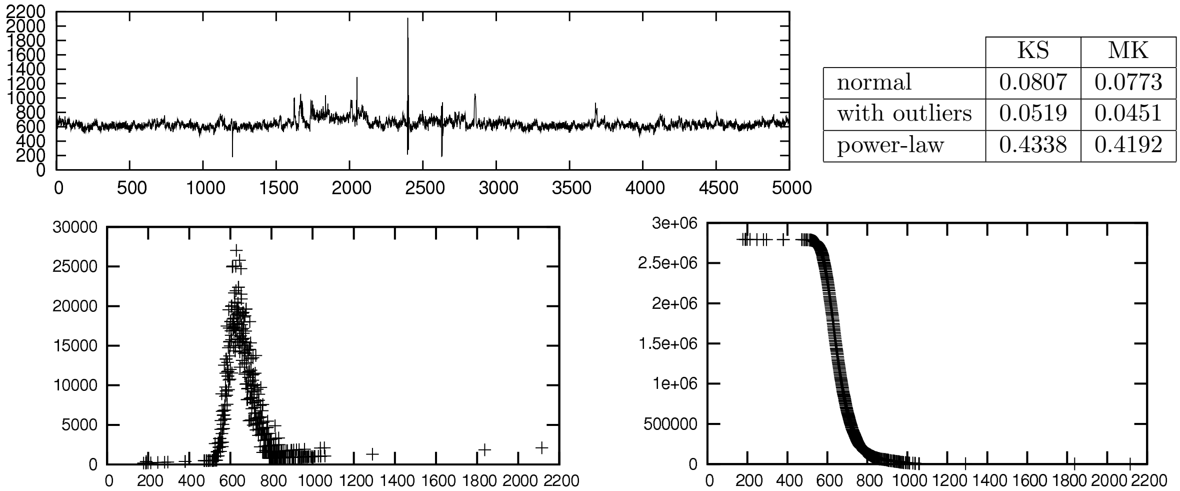
The obtained plot exhibits clear upper peaks, independent of the current mean value of N_i^c, which is confirmed by the distributions and fit tests. We obtain this way a method for automatic detection of statistically meaningful events, defined as outliers in the number of appearing nodes.

5 Connected components

We have seen in the previous section that, at some particular moments, an abnormal number of nodes appear in our ego-centered views of the internet topology. However, we said nothing on their *structure*: are they scattered in the observed topology? are they grouped? or do they belong to several small groups?... Intuitively for instance, an important routing change may lead to the discovery of a new part of the network, which would be revealed by the appearance of nodes forming a connected component in our ego-centered views.

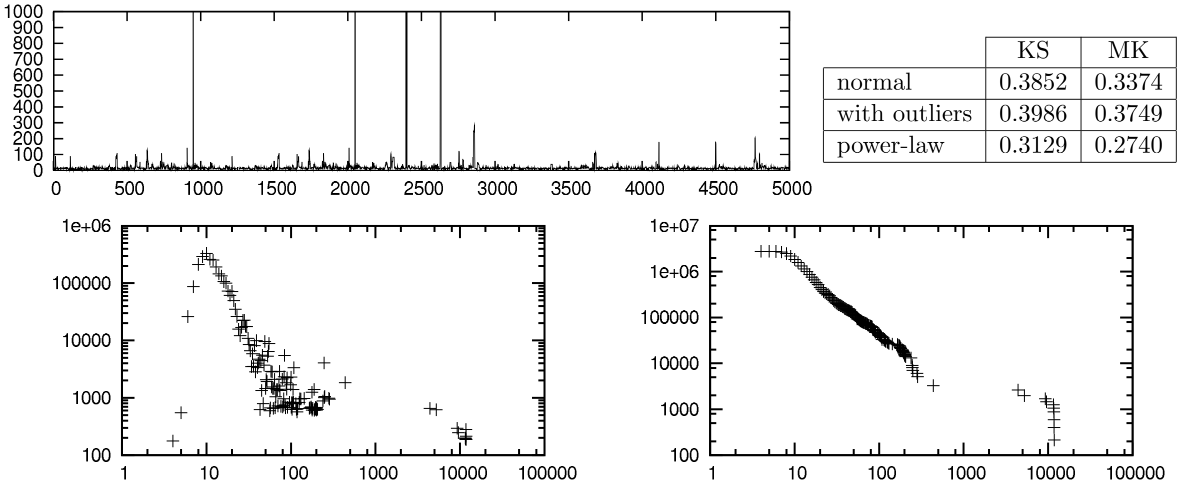
In order to investigate this, we study the connected components of newly appearing nodes. More precisely, for all i, we select the appearing nodes as defined above, and consider links observed between these nodes. We then compute the connected components of this graph, which we call *connected components of appearing nodes*. As in the previous section, we use $p = 10$ and $c = 2$, which give results representative of what we observed on wide ranges of these values.

We show in Figure 9 the number of connected components observed for all i, and the size of the largest one for all i in Figure 10, together with the distributions of these values and their goodness of fit test. The statistically abnormal events detected with the number of connected components statistics are the same as the ones detected in the previous section. This shows that events detected using the number of appearing nodes are events in which many connected components appear, among which at least a large one.



	KS	MK
normal	0.0807	0.0773
with outliers	0.0519	0.0451
power-law	0.4338	0.4192

Figure 9: Number of connected components of nodes appearing at round i, as a function of i, and the distribution and inverse cumulative distribution of this value. The table displays KS and MK tests for the distribution with each considered model distribution. We used representative values $p = 10$ and $c = 2$.

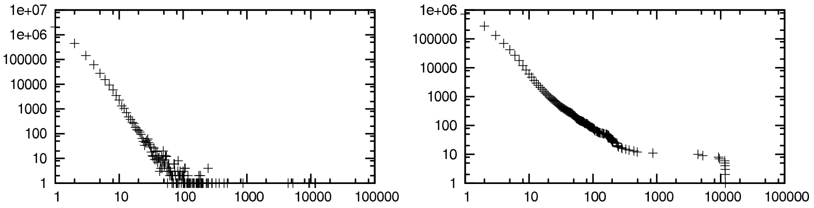


	KS	MK
normal	0.3852	0.3374
with outliers	0.3986	0.3749
power-law	0.3129	0.2740

Figure 10: Size of the largest connected component of nodes appearing at round i, as a function of i, and the distribution and inverse cumulative distribution of this value. The table displays KS and MK tests for the distribution with each considered model distribution. We used representative values $p = 10$ and $c = 2$.

Observing connected components makes it possible to go further. Indeed, it has the advantage that, at each round, several values are observed: for all i several connected components may appear, and we may consider their size. This leads to the distribution of the size of *all* appearing connected components, whichever round they appear in, presented in Figure 11.

This distribution does not exhibit a clear difference between *normal* values and *abnormal* ones, though: the distribution is well fitted by a power-law, as the goodness of fit test in Figure 11 shows. As a consequence, we cannot use it to detect events that would be revealed by the appearance of an abnormally large connected component.



	KS	MK
normal	0.5653	0.5490
with outliers	0.5642	0.5100
power-law	0.1510	0.1407

Figure 11: Distribution and inverse cumulative distribution of the size of *all* connected components of appearing nodes. The table displays KS and MK tests for the distribution with each considered model distribution. We used representative values $p = 10$ and $c = 2$.

Notice that one may go further by computing various properties of connected components (their density, average degree, or clustering coefficient, for instance), and then observing their distribution. This may lead to the identification of statistically meaningful events. However this is out of the scope of this paper and left for future work.

6 Distance-based properties

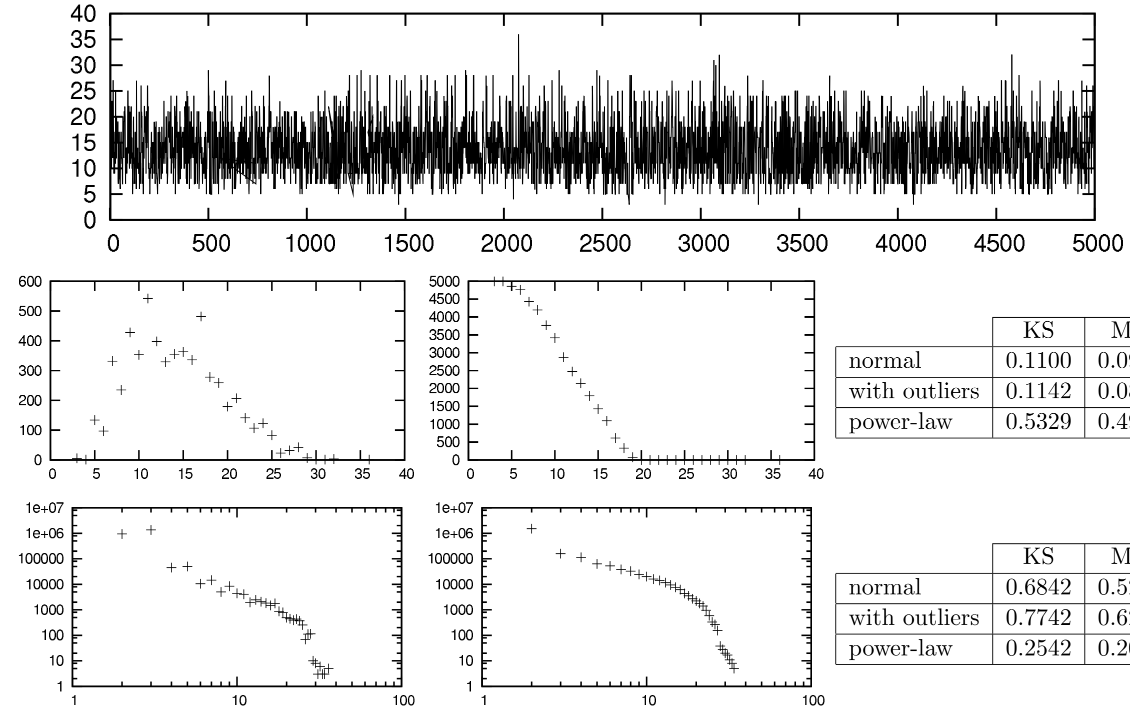
Properties considered in previous sections remain rather basic and only capture poor structural features (size and connectivity). In this section we consider distance-based properties[4] in order to capture more subtle features.

Similarly to appearing nodes considered above, we define for any integer i the *appearing links* as the links observed in any of the i to $i+c-1$ rounds of measurement but in none of the $i-p$ to $i-1$ rounds, for given numbers p of previous rounds and c of current rounds. With our notations, they are therefore links in $E_i^c \setminus E_{i-p}^p$. Notice that all the links of an appearing node necessarily are appearing links; we will call them *trivial* appearing links, and focus on *nontrivial* appearing links. Nontrivial appearing links are therefore links that appear during current rounds between nodes which were present but not linked together in previous rounds: appearing links are the links in $(E_i^c \setminus E_{i-p}^p) \cap (V_{i-p}^p \times V_{i-p}^p)$.

One may expect that nontrivial appearing links tend to appear between nodes which were already close in the previous rounds, *i.e.* such that the distance between them in G_{i-p}^p is small. Notice that there is one value of the distance for each such link, and therefore we have to consider several values at each round (one per nontrivial appearing link). Let us denote by D_i the set of values obtained for round i.

We plot in Figure 12 the value of $\max(D_i)$ as a function of i, as well as the distribution of these values. We also plot the distribution of all observed distances, *i.e.* $\cup_i D_i$. For these values, there is no meaningful time series, but as before the fact that there are several values for each i has the advantage of making results more statistically sound, and one may expect to detect several events occurring at the same time.

[4] The *distance* between two nodes is the length of a shortest path between them, *i.e.* the number of hops needed to go from one to the other in the (undirected) graph.



	KS	MK
normal	0.1100	0.0952
with outliers	0.1142	0.0898
power-law	0.5329	0.4921

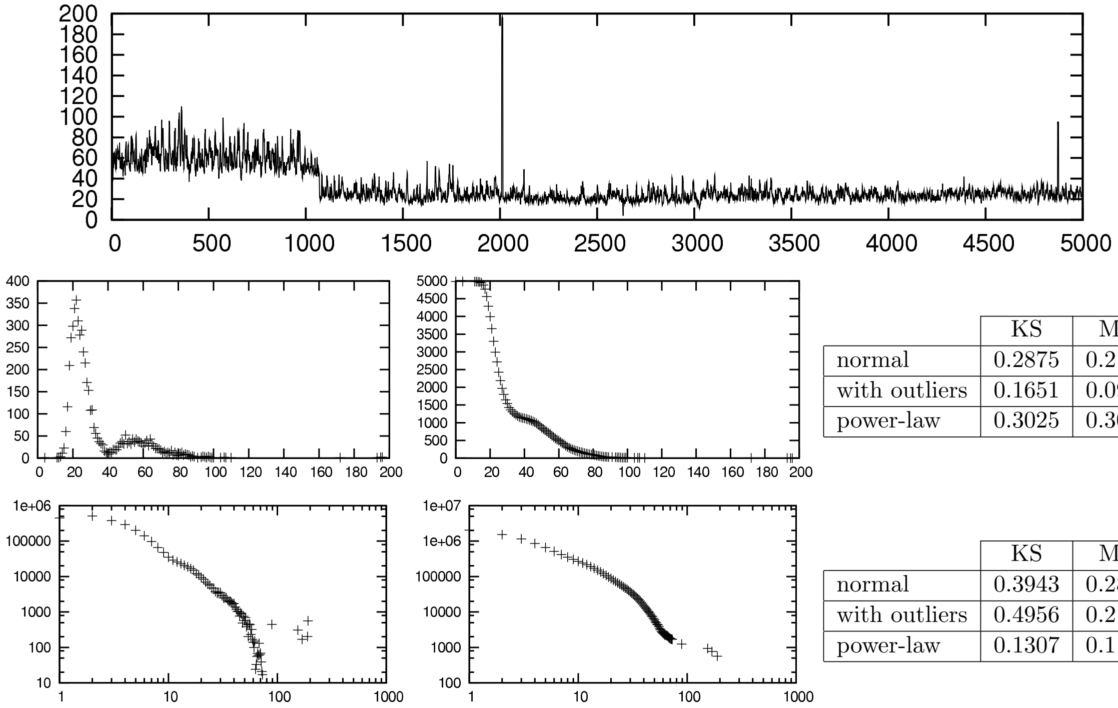
	KS	MK
normal	0.6842	0.5268
with outliers	0.7742	0.6260
power-law	0.2542	0.2007

Figure 12: Top row: largest distance between extremities of nontrivial appearing links, *i.e.* $\max(D_i)$, as a function of i. Second row: distribution and inverse cumulative distribution of these values, and KS and MK tests for this distribution. Third row: distribution and inverse cumulative distribution of all distances between extremities of appearing links, *i.e.* $\cup_i D_i$ (several values per round), and KS and MK tests for this distribution. We used representative values $p = 10$ and $c = 2$.

As expected, the observed distances are rather small, even though relatively large values (up to 35) occur. This may be considered as large as the distances in considered graphs are known to be small (not significantly higher than this extremal value). This being said, the distribution of the maximal distance is homogeneous. Automatic tests show that extremal values may be considered as outliers, but the difference with purely homogeneous distribution is low. If one considers *all* distances, the distribution becomes heterogeneous, although the maximal is by definition identical. The best fit is the power-law one, and so this property cannot be directly used to detect events.

One key problem with these statistics is that all distances in the considered graphs are small, and are very similar among nodes (the distribution of all distances in the graph is homogeneous). As a consequence, there is little hope that distributions of distances between nodes, even selected pairs of nodes, may exhibit distributions rich enough for event detection. In order to obtain such distributions, one has to consider properties with values spanning a wider interval.

In order to find such a property while still relying on distances, we define for any nontrivial new link (u, v) the number of nodes in V^p_{i-p} such that their distance to u or v is not the same in the previous and current graphs, G^p_{i-p} and G^c_i respectively. We denote this number by $\delta(u, v)$. Like above, for each i we define Δ_i as the set of $\delta(u, v)$ for all nontrivial appearing links at round i.



	KS	MK
normal	0.2875	0.2116
with outliers	0.1651	0.0993
power-law	0.3025	0.3002

	KS	MK
normal	0.3943	0.2862
with outliers	0.4956	0.2190
power-law	0.1307	0.1192

Figure 13: Top row: largest number $\delta(u, v)$ of nodes with changing distance to extremities of appearing links at i, *i.e.* $\max(\Delta_i)$, as a function of i. Second row: distribution and inverse cumulative distribution of these values for $i > 1100$ (to avoid the change in regimes which would disrupt the distribution), and KS and MK tests for this distribution. Third row: distribution and inverse cumulative distribution of $\delta(u, v)$ for all appearing links (u, v) at any $i > 1100$, *i.e.* $\cup_{i>1100}\Delta_i$ (several values per round), and KS and MK tests for this distribution. We used representative values $p = 10$ and $c = 2$.

We plot in Figure 13 the value of $\max(\Delta_i)$ as a function of i, as well as the distribution of these values. We also plot the distribution of all obtained values, *i.e.* $\cup_i\Delta_i$.

As expected, the values of this property span a significantly wider range than simple distances. More importantly, it succeeds in exhibiting two kinds of events: the maximal value for each i oscillates close to an average value but exhibits abnormally high values as well as changes in the average value. Both phenomena are clearly visible in Figure 13, and confirmed by the distributions. Both provide a way to automatically detect events with a distance-related property.

The situation is not as clear when one considers *all* values: Figure 13 shows that the distribution is rather heterogeneous, even though visual inspection may indicate that values over 100 constitute events. Still, our automatic tests consider the distribution as heterogeneous and misses these events. Certainly, more subtle statistical techniques could be used to improve this, but this is out the scope of this paper.

7 Correlations between detected events

In previous sections, we have defined and studied various properties aimed at detecting events in the dynamics of a graph, here ego-centered measurements of the internet topology. Several led to homogeneous distributions with outliers, and thus are effective for this purpose. One may however wonder if all properties detect the same events (in which case subtle and more costly properties would not be useful) or if they detect different events (in which case they are all useful and complementary). In order to explore this, we study here correlations between events detected by each property.

To do so, we plot together in Figure 14 the three main properties that proved to be relevant for event detection: the variations in the median of the number of nodes in each round (Section 4.1); the maximal number of nodes with changing distance due to an appearing link (Section 6); and the number of appearing nodes (Section 4.2).

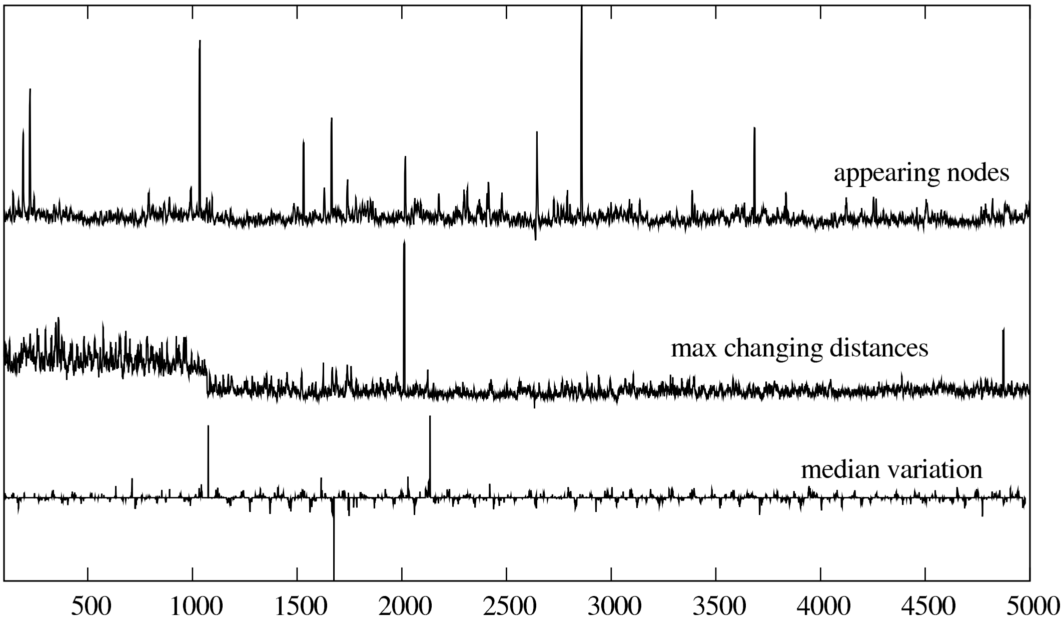


Figure 14: The three main properties that proved to be relevant for event detection. From bottom to top: the variations in the median of the number of nodes in each round (Section 4.1); the maximal number of nodes with changing distance due to an appearing link (Section 6); and the number of appearing nodes (Section 4.2). As previously, we used $p = 10$ and $c = 2$. We scaled and shifted plots for readability (values on the vertical axis would have no meaning so we do not display any).

Several important observations may be made from this figure. First, for each property, there exists an event (a peak in the plots) which is detected by this property only. For instance, there is a clear event just before the 5000-th round of measurement which is detected by the distance-based property only. This shows that all discussed graph properties make sense regarding event detection and should be considered as complementary. Instead, some events are detected by several, and even all, statistics, like the one slightly after $i = 1000$. This shows that some events have an impact on several properties while others do not, and thus that our approach makes a difference between different classes of events.

Finally, correlations between detected events are nontrivial. Basically all possible situations occur, which shows that the set of detected events probably is very complex and rich (although limited in size, which is an important feature). We discuss some methods to study it in the following section, but fully exploring this direction remains one of our main perspectives.

8 Towards event interpretation

In the previous sections we have presented a methodology and some statistics which make it possible to detect statistically significant events. More precisely, we are able to point out moments in time at which events occur, and to identify nodes and links involved in these events. The ultimate goal of this procedure is to study detected events further and in particular to *interpret* them in term of network events (such as node or link failures, or congestions). This is crucial for a true understanding of internet dynamics and for network monitoring.

Event interpretation is however challenging, because the current knowledge of the internet dynamics is limited, but also because of the size of the data, its ego-centered (thus biased) nature, and its lack of clear structure. Ideally, one may use a database of events occurring in the internet and match such events to the ones we detect (and conversely). This is not feasible in general, though, as no complete such database exists. Only partial information is available for some specific AS, which we explore in Section 8.1 below.

One may also try to interpret detected events by visualizing the data. To do so, graph drawing is appealing, but current methods are unable to handle large graphs and/or produce drawings which are easy to interpret. Some insight may however be obtained in this way, and we explore this in Section 8.2.

In the following, we select one of the most interesting statistics for detecting events, the number of appearing nodes in several consecutive rounds (Section 4.2). We apply it to a typical measurement and select events detected in this way.

Moreover, we use a data reduction technique which is of great help. It consists in focusing on the part of the data involved in the event under concern. To do so we first identify the set S of nodes involved in the event (this depends on the considered property). We then select the destinations such that a path from the monitor to the destination contains at least a node in S. Finally we keep only the part of the measurement obtained with these destinations, which is equivalent to measurements conduced with this reduced set of destinations. After this reduction, all data involving nodes in S is still present and so we expect to capture most of the dynamics related to the event under concern. Instead, we expect to remove much data about nodes not involved in this event, which helps much in studying it (visualisation tasks in particular).

8.1 Correlations with known events

In order to help maintenance and provide better services, some ISP record *events* occurring in their network and document them. This information is partial, poorly structured, and needs manual inspection [23, 19], but this is of great of interest here as it makes it possible to match statistically significant events which we detect to known network events reported in these databases.

Abilene [1] is one the main ISP to provide rich information on events occurring in their network [42]. A database of tickets describing such events is freely available online; we display typical instances in Figure 16.

To match a detected event to such a ticket, we proceed as follows. First we select a statistically meaningful event with our method as explained above, and then we localize the

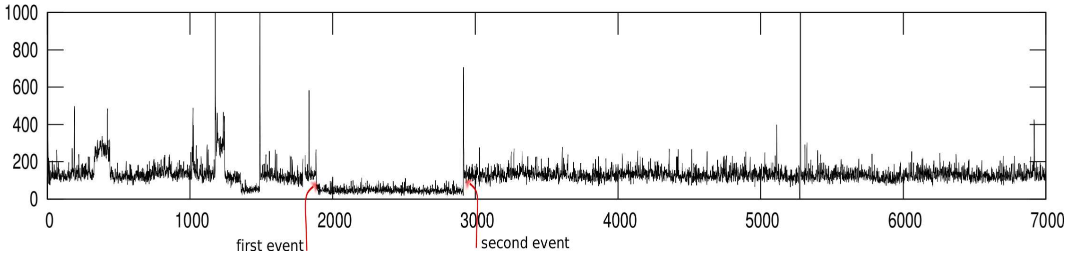


Figure 15: Number a_i of appearing nodes, as a function of the number of measurements rounds performed. The two arrows denote two statistical events that were correlated with two known events tracked by the Abilene trouble tickets of Figure 16. Here we considered $p = 10$ and $c = 2$.

timestamps at which it happens (which correspond to peaks on the corresponding plot). Correlating this event with an Abilene event consists in finding in the Abilene database a set of tickets such that the timestamps of these tickets overlap the timestamps of our event, and the *affected* fields cite elements with addresses belonging to the set S of involved nodes. We therefore have to collect the IP addresses of the elements cited in the ticket in order to check their presence in S.

An example of result is displayed in Figure 15. Among the detected events, two of them are correlated with tickets. The first statistical event that we pointed out is followed by a significant decrease in the number of appearing nodes. This event is correlated with the Abilene trouble ticket shown in Figure 16 (left). A second event occurs later, followed by an equivalent significant increase in the number of appearing nodes. Inspecting this second event leads to its correlation with the other trouble ticket in Figure 16, which turns out to be the ticket declaring that the problem cited in the first ticket ended. In this case, thus, there is a perfect fit between the two statistical events under concern and the one depicted in Figure 16.

```
SUBJECT:       Internet2 IP Network Peer SINET (CHIC) Outage          | SUBJECT:       Internet2 IP Network Peer SINET (CHIC) Resolved
AFFECTED:      Peer SINET (CHIC)                                      | AFFECTED:      Peer SINET (CHIC)
STATUS:        Unavailable                                            | STATUS:        Available
START TIME:    Thursday, May 17, 2007, 11:47 AM (1147) UTC            | START TIME:    Thursday, May 17, 2007, 11:47 AM (1147) UTC
END TIME:      Pending                                                | END TIME:      Friday, May 18, 2007, 3:51 AM (0351) UTC
DESCRIPTION:   Peer SINET's connection the Internet2 IP               | DESCRIPTION:   Peer SINET was unavailable to the Internet2 IP
               Community is unavailable.  SINET Engineers             |                Network Community.  SINET Engineers reported the
               have been contacted, however, no cause of              |                reason for outage was due to a fiber cut in New York.
               outage has been provided yet. SINET is multi-homed.    |                SINET is multi-homed.
TICKET NO.:    10201:45                                               | TICKET NO.:    10201:45
TIMESTAMP:     07-05-18 00:40:43 UTC                                  | TIMESTAMP:     07-05-18 07:39:16 UTC
```

Figure 16: Examples of Abilene trouble tickets which corresponds to the events pointed out in Figure 15. Left, corresponding to the first event, describes a technical intervention under the ticket number 10201:45. The involved network elements are cited in the field AFFECTED. The begin and the end timestamps are given, and details are provided in field DESCRIPTION. Right corresponds to the second event.

8.2 Graph drawing

One may also examine a detected event by manipulating a drawing of the underlying graph. Although many drawing methods exist, with different advantages and limitations, in most

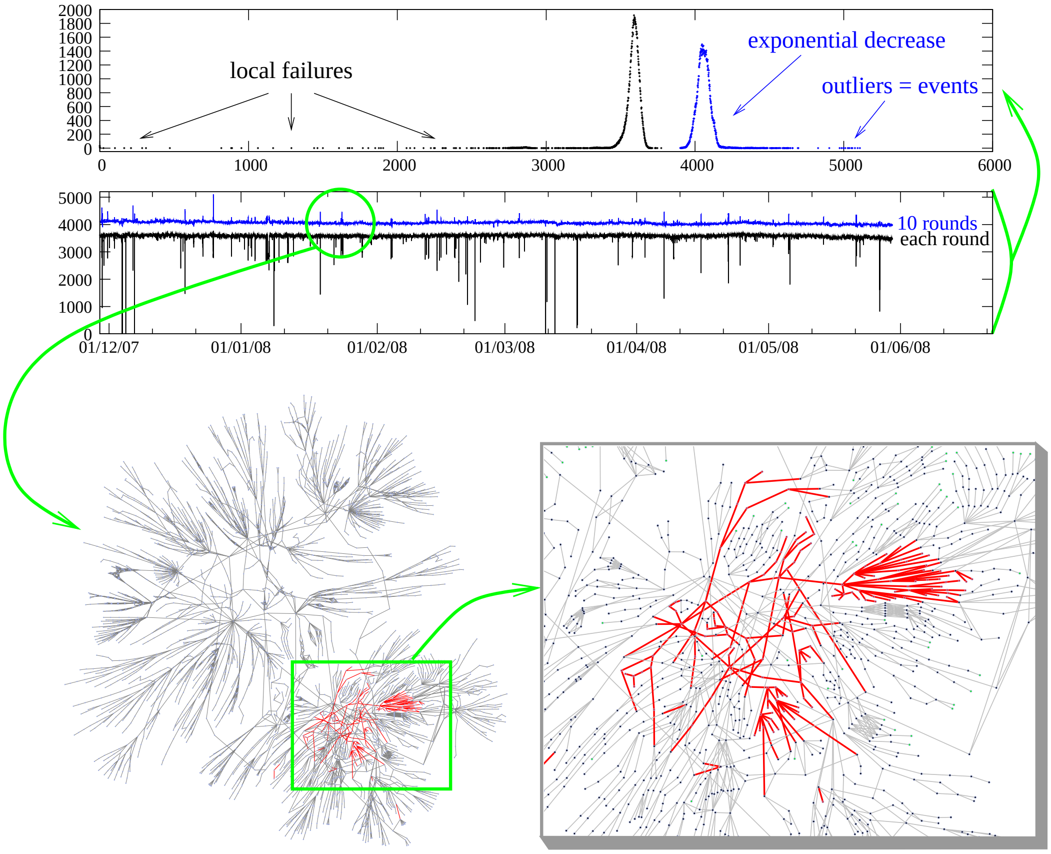


Figure 17: Middle row: plot of the number of distinct nodes N_i (resp. N_i^{10}) observed during each round (resp. ten consecutive rounds) of measurement, as a function of time. Top row: the distributions of these values, which confirms that N_i exhibits abnormally small values only, never abnormally large ones, unlike N_i^{10}. Bottom row: topology changes observed during an event identified by an abnormally large value of N_i^{10}, with a zoom on the part of the network where the event occurred. Appearing nodes and links are in red.

cases the size of our data is prohibitive. To this regard, being able to identify a moment in time at which an event occurs and focusing on the nodes involved in the event as described above are both crucial: this data reduction leads to graphs of a few thousand nodes, which several software are able to manipulate (and draw).

One may then draw in different colors the appearing, disappearing and stable nodes and/or links. Figure 17 displays a typical example. We observe that, whereas the dynamics is in general scattered in all the network, this event corresponds to a significant change in a specific part of the topology.

Such manual examination of events using graph manipulation software opens the way to a more detailed understanding of detected events, and to their interpretation in terms of network events.

9 Related work

Our work is at the intersection of two topics: event detection and dynamics of internet topology (and dynamic graphs in general).

With regard to event detection, this is, up to our knowledge, the first work to consider *graph dynamics*. This led us to introduce properties of such dynamics, from trivial ones like the number of nodes at each time step to more subtle ones like the distance-based ones with several (and a variable number of) values at each time step. Notice moreover that our method is very general and may be applied directly to many other dynamic graphs (social networks, for instance), which is an important contribution in itself.

The problem of event detection however is far from being new. This is a classical problem in many contexts, including the internet. In particular, much attention has been paid to event detection in internet traffic, see for instance [4, 26, 40, 22, 28, 14, 7]. In such cases, though, the analysis mostly considers time series (*i.e.* one value for each time unit) which represent simple quantities such as the number of packets captured within the measurement. Subtle structural properties like the ones considered here are not used. Notice however that here too we often deal with time series, and that subtle methods developed for time series analysis may be applied. This is an important perspective of our work.

More generally, many studies target event detection in the dynamics of various systems [9]. Two main approaches are followed, named anomaly-based and signature-based.

The underlying principle of anomaly-based approaches [6, 5] is that one knows the normal behavior of the system. Then, any observation that differs from this normal behavior is considered as an event. This approach is very appealing as it is able to detect any kind of event, including kinds that were never observed. It however relies on a precise knowledge of the dynamics of the considered object, and evolution of the normal behavior makes the method ineffective. In the case of the internet topology, these two limitations make this approach unapplicable.

Signature-based approaches rely on the knowledge of characteristic features of events to detect, which may be inferred from a set of known events (typically with machine learning techniques) [35, 9]. If the observed dynamics matches these features at some point, then one considers that this is an event. This approach is very effective in cases where the events may be described, like some computer viruses for instance. In our case, though, very limited knowledge of events in the internet, and no description on their impact on observed topology, are available.

As a consequence, both classical approaches are unapplicable in our context. This is why we developed a statistically based approach, which may miss interesting events and calls for the definition of relevant statistics (which may be challenging), but does not require any previous knowledge of the system and event features. This makes the method very general and robust. The main drawback is that detected events may be difficult to interpret.

Notice that, as intuition on the dynamics of internet topology is often misleading (see for instance [30]), and as it is not known what characterizes normal dynamics in this context [8], our method may be seen as a way to dig into measurement data. It provides insight on what may be seen as normal dynamics and what indicates events. It points out specific moments in time when something unusual happens and identifies sets of involved nodes and links. This makes it possible to investigate further the dynamics, thus providing one of the most efficient tool currently available for the empirical study of graph dynamics.

Finally, let us insist on the fact that the dynamics of the internet topology is at the center of much interest. Still a limited number of previous works study it, though. For instance, [3, 37] study the dynamics induced by load balancing; [36, 24, 25] study the dynamics of BGP routing; [33] studies the long-term evolution of the AS-level topology; [30] studies specific features of the dynamics of ego-centered views like the ones considered here; [34] studies the dynamics of the multicast routers topology; etc. None of these works target event detection, though, and they generally consider dynamics at a much coarser grain than us (days vs minutes).

10 Conclusion

In this paper we propose and implement a method to automatically and rigorously detect events in the dynamics of ego-centered views of the internet topology. It relies on a notion of statistically significant events. We define statistics to do so, some simple and others more subtle. Interestingly, all kinds of distributions are obtained: homogeneous and heterogeneous, which do not lead to the detection of events, and homogeneous with outliers, which do. In addition, we show that different properties lead to the discovery of different events, and therefore that it makes sense to try and define more dynamic graph properties to gain more insight. We also provide approaches to interpret detected events by drawing them and comparing them to known networking events.

A natural perspective for this work is to explore more subtle statistics. In particular, as the data we consider here are ego-centered views of a network, using statistics designed specifically for such situations would be very interesting, see for instance [13]. One may also observe events using measurements conducted from several monitors. As some events may be invisible from single monitors, this approach is very promising. One may detect events on views obtained by merging measurements from several monitors, or detect events from each measurement independently and then mix the observations.

Another key direction is to conduct more studies of detected events to gain insight on the underlying causes and their effect. In order to help such interpretation, one may simulate ego-centered measurements on a graph with simulated dynamics (random removals/additions of nodes/links, for instance). This would shed light on the relation between what we observe with such measurements and actual events on the topology, which is crucial in our context. Designing appropriate models for doing so however is challenging [29, 31], and most remains to be done in this direction.

Finally, as our method is very general, an appealing direction is to apply it to other case studies (like social networks, for instance). In such cases, event interpretation may be easier. We expect to obtain in this way the first method for empirical analysis of graph dynamics, and in particular unusual events in such dynamics.

Acknowledgements

This work is partly funded by the European Commission through the EULER project (grant 258307), part of the *Future Internet Research and Experimentation (FIRE)* objective of the Seventh Framework Programme (FP7). It is also partly funded by the *DynGraph* grant from the *Agence Nationale de la Recherche* with reference ANR-10-JCJC-0202.

References

[1] Abilene backbone network. `http://www.internet2.edu/`.

[2] Caida – Archipelago project. `http://www.caida.org/projects/ark/`.

[3] Brice Augustin, Xavier Cuvellier, Benjamin Orgogozo, Fabien Viger, Timur Friedman, Matthieu Latapy, Clémence Magnien, and Renata Teixeira. Traceroute anomalies: Detection and prevention in internet graphs. *Computer Networks*, 52, 2008.

[4] P. Barford, J. Kline, D. Plonka, and A. Ron. A signal analysis of network traffic anomalies. In *ACM Internet Measurement Workshop*, 2002.

[5] Damiano Bolzoni and Sandro Etalle. Approaches in anomaly-based intrusion detection systems. In *1st Benelux Workshop on Information and System Security*, 2006.

[6] Damiano Bolzoni, Sandro Etalle, and Pieter Hartel. Poseidon: a 2-tier anomaly-based network intrusion detection system. In *Fourth IEEE International Workshop on Information Assurance (IWIA)*, 2006.

[7] P. Borgnat, G. Dewaele, K. Fukuda, P. Abry, and K. Cho. Seven Years and One Day: Sketching the Evolution of Internet Traffic. In *Proceedings of the 28th IEEE INFOCOM 2009*. IEEE, 2009.

[8] Jake D. Brutlag. Aberrant behavior detection in time series for network monitoring. In *LISA '00: Proceedings of the 14th USENIX conference on System administration*, pages 139–146, Berkeley, CA, USA, 2000. USENIX Association.

[9] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3), 2009.

[10] Qian Chen, Hyunseok Chang, Ramesh Govindan, Sugih Jamin, Scott J. Shenker, and Walter Willinger. The origin of power laws in internet topologies revisited, 2002.

[11] A. Clauset, C.R. Shalizi, and M.E.J. Newman. Power-law distributions in empirical data. *SIAM Review*, 4(51):661–703, 2009.

[12] M.J. Crowder. Parameter estimation for scientists and engineers by Adriaan van den Bos. *International Statistical Review*, 75(3):436–437, December 2007.

[13] L. da F. Costa and R.F.S. Andrade. What are the best concentric descriptors for complex networks? *New Journal of Physics*, 9:311, 2007.

[14] G. Dewaele, K. Fukuda, P. Borgnat, P. Abry, and K. Cho. Extracting hidden anomalies using sketch and non gaussian multiresolution statistical detection procedures. In *SIGCOMM 2007 Workshop LSAD - ACM SIGCOMM 2007 Workshop on Large-Scale Attack Defense (LSAD)*, 2007.

[15] Dimes website. `http://www.netdimes.org/new/`.

[16] Scott R Eliason and Michael S Lewis Beck. *Maximum Likelihood Estimation: Logic and Practice*. Sage Publications (CA), 1993.

[17] Tryphon T. Georgiou, Johan Karlsson, and Mir Shahrouz Takyar. Metrics for power spectra: An axiomatic approach. *IEEE Transactions on Signal Processing*, 57(3):859–867, 2009.

[18] D. Hawkins. *Testing for Normality.* Springer, 1980.

[19] Yiyi Huang, Nick Feamster, and Renata Teixeira. Practical issues with using network tomography for fault diagnosis. *Computer Communication Review 38(5)*, pages 53–58, 2008.

[20] University of washington – iPlane project. `http://iplane.cs.washington.edu/data.html`.

[21] V. Jacobson. traceroute, February 1989. The most recent version is available at: `ftp://ftp.ee.lbl.gov/traceroute.tar.gz`.

[22] B. Krishnamurty, S. Sen, Y. Zhang, and Y. Chen. Sketch-based Change Detection: Methods, Evaluation, and Applications. In *Proceedings of ACM IMC*, Miami, 2003.

[23] Amelie Medem Kuatse, Renata Teixeira, and Mickael Meulle. Characterizing network events and their impact on routing. *CoNEXT*, page 59, 2007.

[24] C. Labovitz, G. Robert Malan, and F. Jahanian. Origins of internet routing instability. In *Proc. IEEE INFOCOM*, pages 218–226, 1999.

[25] M. Lad, D. Massey, and L. Zhang. Visualizing internet routing changes. *IEEE Transactions on Visualization and Computer Graphics, special issue on Visual Analytics*, 2006.

[26] A. Lakhina, M. Crovella, and C. Diot. Diagnosing Network-Wide Traffic Anomalies. In *Proceedings of ACM SIGCOMM '04*, 2004.

[27] Matthieu Latapy, Clémence Magnien, and Frédéric Ouédraogo. A radar for the internet. *Complex Systems*, 20(1), 2009.

[28] X. Li, Bian. F., M. Crovella, C. Diot, R. Govindan, A. Lakhina, and G. Iannaccone. Detection and Identification of Network Anomalies Using Sketch Subspaces. In *Proceedings of ACM IMC*, Rio de Janeiro, 2006.

[29] Clémence Magnien, Amélie Medem, Sergey Kirgizov, and Fabien Tarissan. Towards realistic modeling of ip-level routing topology dynamics. *Networking Science*, 2013. To appear.

[30] Clémence Magnien, Frédéric Ouédraogo, Guillaume Valadon, and Matthieu Latapy. Fast dynamics in internet topology: preliminary observations and explanations. *Fourth International Conference on Internet Monitoring and Protection (ICIMP 2009)*, 2009.

[31] Amélie Medem, Clémence Magnien, and Fabien Tarissan. Impact of power-law topology on ip-level routing dynamics: simulation results. *NetSciCom*, 2012.

[32] T. Moors. Streamlining traceroute by estimating path lengths. In *Proc. IEEE Workshop on IP Operations and Management*, October 2004.

[33] Ricardo V. Oliveira, Beichuan Zhang, and Lixia Zhang. Observing the evolution of internet AS topology. *SIGCOMM Comput. Commun. Rev.*, 37(4):313–324, 2007.

[34] J.-J. Pansiot. Local and dynamic analysis of internet multicast router topology. *Annales des télécommunications*, 62:408–425, 2007.

[35] Janak J. Parekh, Ke Wang, and Salvatore J. Stolfo. Privacy-preserving payload-based correlation for accurate malicious traffic detection. In *Proceedings of the 2006 SIGCOMM workshop on Large-scale attack defense*, LSAD '06, pages 99–106, New York, NY, USA, 2006. ACM.

[36] S.-T. Park, D. M. Pennock, and C. L. Giles. Comparing static and dynamic measurements and models of the internet's AS topology. In *Proc. IEEE Infocom*, 2004.

[37] V. Paxson. End-to-end internet packet dynamics. *IEEE/ACM Trans. Networking*, 7(3):277–292, June 1999.

[38] William H. Press, Saul A. Teukolsky, William T. Vetterling, and B. P. Flannery. *Numerical Recipes: The Art of Scientific Computing.* Cambridge University Press, Cambridge, England, 1992.

[39] Radar datasets. At `http://data.complexnetworks.fr/Radar/`.

[40] A. Scherrer, N. Larrieu, P. Owezarski, P. Borgnat, and P. Abry. Non gaussian and long memory statistical characterisations for internet traffic with anomalies. *IEEE Trans. on Depend. and Secure Comp.*, 4(1):56–70, January 2007.

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

[42] Technical discussion for the internet2 network. At `https://listserv.indiana.edu/archives/internet2-ops-l.html`.

[43] Henry C. Jr. Thode. *Identification of Outliers.* CRC Press, 2002.

[44] Traceroute@home website. `http://www.tracerouteathome.net/`.

[45] Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference.* Springer, 2005.