

# Estimating the access link quality by active measurements

Roberto G. Cascella

INRIA Sophia Antipolis, EPI Pleanete  
2004, rue des Luciolles  
Sophia Antipolis, France  
Email: Roberto.Cascella@sophia.inria.fr

Chadi Barakat

INRIA Sophia Antipolis, EPI Pleanete  
2004, rue des Luciolles  
Sophia Antipolis, France  
Email: Chadi.Barakat@sophia.inria.fr

**Abstract**—The access link quality experienced by the end users depends on the amount of traffic and on the presence of network anomalies. Different techniques exist to detect anomalies, but little attention has been devoted to quantify the access link quality and to which extent network anomalies affect the end user’s access link experience. We refer to this aspect as the *impact factor* of the anomaly, that we define as the percentage of affected destinations. In the ideal case, a node should continuously monitor all possible routes to detect any degradation in performance, but this is not practical in reality.

In this paper we show how a node can estimate the quality of Internet access through a limited set of measurements. We initially study the user’s access network to understand the typical features of its connectivity tree. Then, we define an unbiased estimator for the quality of access and we compute the minimum number of paths to monitor, so that the estimator achieves a desirable accuracy without knowing the underlying topology. We use real data to construct a network graph and we validate our solution by causing a large number of anomalies and by comparing the real and the estimated quality of access for all available end hosts. Our results show that the impact factor is a meaningful metric to evaluate the quality of Internet access.

## I. INTRODUCTION

The rapid expansion of applications and services, e.g., voice over IP and video on demand, needs to face with the existent Internet architecture controlled by Internet Service Providers (ISPs). In this context, ISPs attract users offering flat rate subscriptions with a minimum guaranteed bandwidth for their traffic and rely on overprovisioning to meet their application requirements. However, the quality of service perceived by the user depends on many factors which might be out of the control of the ISP: link failures and router congestion inside the ISP network, or in other transit networks along the path, can lead to a service degradation; saturation of servers involved in the communication also impacts end user performances. The detection of these network anomalies is important for the ISP to manage the state of the traffic and the network, but it is also very important for the end user to assess the quality of its connection. In absence of strict guarantees on the quality of service by ISPs, customers are left themselves to judge the perceived quality of service and estimate whether the ISP is fulfilling the Service Level

Agreement (SLA). In the ideal case, the customers should continuously monitor their connection and use tools, like those provided by Grenouille [1], which allow measuring a set of metrics, such as the speed of the downlink and uplink of the end user’s connection. However, ISPs are hostile to such type of measurements and often do not collaborate by blocking measurement traffic or in the worst case by rerouting the users’ traffic so that the problem looks to be caused by other ISPs.

Generally, the detection of network anomalies and the assessment of the Internet access by end users are quite challenging due to the limited information available to them and the limited set of achievable measurements. In recent years, researchers have focused on estimating a set of metrics, such as packet loss, bottleneck bandwidth, and round trip time (RTT), and numerous solutions have been proposed to detect anomalies [2]–[6]. The main idea is to infer network anomalies by variation of path properties based on past observed events. Network tomography is also a well established research subject that uses end-to-end measurements to identify internal characteristics of the network without any cooperation from network devices [7], [8], and it has been used for anomaly detection with topology information [9]. Although the problem of detecting network anomalies has received a considerable amount of attention, little has been done in estimating the real impact of these anomalies on the user’s Internet access and in determining scalable approaches that can work at the edges to assess the quality of the network in general.

In this paper we take a first step toward the network anomalies’ characterization at the edges of the network. We define an anomaly as a deviation from the normal function of the network due to a change in the path metrics, such as end-to-end delay, bandwidth and so on. The approach we take is based on inferring general properties from a small set of end-to-end measurements to random destinations, i.e., landmarks, without any assumption on the network topology. The advantage of this approach has been demonstrated for delay estimation using a subset of any network nodes [10] or for end-to-end network monitoring using a subset of paths [11]. We leverage these results to build a model for the characterization of anomalies and for the evaluation of the Internet access provided by ISPs.

This work is motivated by the idea that the end user is primarily interested in knowing what are the consequences

This work was supported by the French ANR C’MON project on Collaborative Monitoring.

of an anomaly in the network more than in a macroscopic diagnosis of the causes of the anomaly. Thus, the question we try to solve in this paper becomes whether there is a network anomaly and if so, what is the impact of this anomaly on the service perceived by the user.

We define the impact factor, or *seriousness*, of an anomaly as being the fraction of destinations and servers, across the Internet, experiencing service degradation. The impact factor ranges between 0 and 1, with 0 being the normal case and 1 the case when all possible IP destinations are affected by the anomaly; the quality of access is no other than one minus the impact factor. As an example, in case of a networking problem, such as congestion in the backbone, the user measures normal performance for local connections but he might have limited access to any service located in other ISPs and reached via the congested backbone link. This anomaly has different impact than similar problems at the access link of the user which will prohibit him from accessing any Internet service. When the problem comes from the server delivering a service to the user, the impact of the anomaly on the user's Internet access is almost nonexistent and the network service can be considered as normal. As a general rule we can reason that the traffic of the user is impacted by an anomaly if the packets traverse the anomalous link and if the destinations reached via this anomalous link are unreachable (or service unacceptable).

Our contributions in this work can be summarized as follows. First we develop a probabilistic model to estimate the value of the impact factor and the minimum number of destinations, i.e., landmarks, to be monitored such that its estimation has a satisfactory accuracy. Second, we use real data to validate in practice our findings and to calculate the number of landmarks, randomly chosen over the set of destinations, so that the estimator of the impact factor achieves a given significance level, without making any assumption on the Internet topology.

The remaining of the paper is organized as follows. The next section motivates our work by analyzing the topology of the access network and by characterizing the properties of the paths to different destinations. Section III introduces the probabilistic model which is used as basis throughout the study. Section IV details the methodology we follow to estimate the impact factor. Section V presents the experimental results for computing the impact factor, the minimum number of landmarks, and the error of the estimation. Section VI discusses the issues associated with the estimation of the impact factor. Section VII concludes the paper and draws future perspectives for this research.

## II. MOTIVATIONS

Network anomalies can be classified based on the events that generate them or on the type of traffic variations they cause in the network [4]. In this paper, we do not specifically study how the network adapts in case of anomalies, i.e., how the routing algorithms compensate for any possible network problem. We rather focus on the consequences of an anomalous event on the users' traffic and on the changes of the quality of

Internet access. An important initial step in this direction is the analysis of the connectivity graph to shed light on the specific features of the users' access network. The user, i.e., the vantage point for monitoring the status of the local connection, can determine the graph using tools such as traceroute.

Let  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  be the connected tree representing the topology having the vantage point as root, nodes  $\mathcal{V}$  as intermediate networking devices, and links  $\mathcal{E}$ . Let  $\mathcal{D}$  be the leaf nodes of the tree, i.e., the destinations of the vantage point's traffic. We now construct the routing matrix  $G \in \{0, 1\}^{|\mathcal{E}| \times |\mathcal{D}|}$  whose entries are  $G_{ij}=1$  if the path to destination  $j$  traverses link  $i$ , otherwise  $G_{ij}=0$ . The paths to the destinations are partially disjoint, which is always the case when destinations do not belong to the same LAN, and these paths share many common links that are close to the vantage point. Thus, the paths are linearly independent and the routing matrix is of full rank. Few dimensions (few paths to the destinations) are more significant than others, such that they might be sufficient to provide a good approximation of the space defined by  $G$ . As a result, the measurements on a subset of paths can be used to estimate the properties and metrics of other paths by leveraging the link sharing property in the Internet. This is the keystone of the tomography problem that deals with the estimation of the internal characteristics of the network [7], [8].

Following these intuitions, we analyze the vantage point connectivity tree on the iPlane data [12] (Section IV-A describes in detail the data). The routing matrix is computed for a total of 137 vantage points from one day's traceroute measurements. There are on average 18,000 destinations per vantage point, with IP addresses in separate /20 networks, and we count on average more than 75,000 unique links per vantage point.

We compute the degree of the connectivity tree and show its distribution versus the number of hops in Fig. 1, where the error bars indicate the standard deviation, and the minimum and the maximum values. Fig. 1 (a) shows the degree distribution of all 137 vantage points and Fig. 1 (b)-(d) show the degree distribution for 3 representative vantage points. It is interesting to notice that in general the degree of the connectivity tree increases at the first few hops and then it falls below 2 after 12 hops. This is explained if we think that the traffic from a vantage point to the destinations first travels along a limited set of paths up to the core where most of the path diversity originates, then uses a limited set of paths to reach destinations in the same ISP. We expect a large set of shared links among paths and the first hop links are shared more than the others.

We then study the spectrum of the tree to illustrate numerically the topological properties of the connectivity trees and this redundancy between paths. We estimate the magnitude of the largest eigenvalues of the routing matrix  $G$  [13] [14] to understand whether few paths are sufficient to represent  $G$  [11]. The analysis of the eigenvalues also gives an insight on the connectivity and clustering properties of the tree. The computation of the eigenvalues is done on  $G \cdot G^T$ , and they are no other than the square of the singular values of the routing matrix  $G$ . We use the Matlab numerical Singular Value

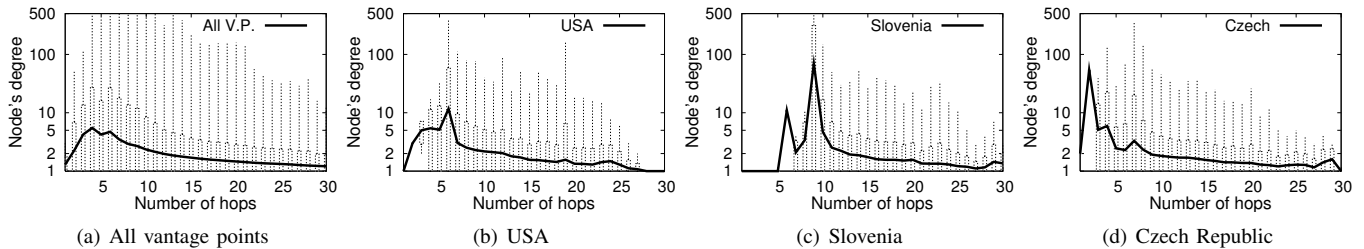


Fig. 1. Degree distribution of the vantage points versus number of hops on semi-logarithmic scale.

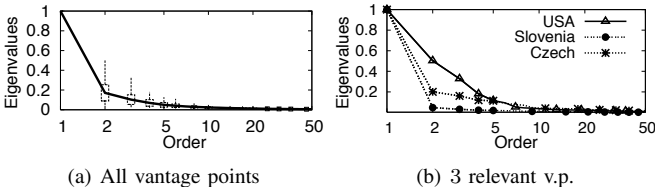


Fig. 2. Largest 50 eigenvalues of the routing matrix of all 137 vantage points (a) and 3 representative vantage points (b).

Decomposition technique for sparse matrices to compute the first 250 singular values.

Fig. 2 (a) plots the mean of the ordered eigenvalues normalized by the largest one for all 137 vantage points. This mean is plotted on a semi-logarithmic scale and the error bars in the plot indicate the standard deviation, the minimum and the maximum values. The sharp decay of the curve indicates that the rank of the routing matrix can be approximated in reality by a smaller value. This ensures that fewer paths cover most of the common initial parts of the tree. Indeed, the rightmost part of the curve shows that the magnitude of the eigenvalues is already negligible after 20. It is worth noticing that the second eigenvalue in Fig. 2 (a) spans from 0.05 to 0.5: this means that the trees have different structures reflecting how the number of links grows with the number of hops [13] [14].

Fig. 2 (b) shows the eigenvalues of the 3 representative vantage points that are geographically distant and exhibit peculiar topology structures. The vantage point located in USA has the largest second eigenvalue compared to others (see Fig. 2 (b)) which characterizes the depth of the tree and indicates that the degrees of the intermediate nodes are comparable. Fig. 1 (b) shows for the same vantage point that the degree distribution increases homogeneously and has a small spike at hop 6. It is worth to analyze the differences with the vantage point located in Czech Republic to derive conclusions on the structure of the tree. Fig. 1 (d) shows a huge spike at the two first hops, which corresponds to a larger number of links at the first hops than the one in USA and in this case the second eigenvalue of the routing matrix is smaller. The different characteristics of the trees should be reflected in the reachability of the destinations which depends on the location of the anomaly and the structure of the tree. We will discuss more about these properties in Section V. The vantage point located in Slovenia shows the opposite extreme case: all the paths share the same links in the first 5 hops, thus the degree of the intermediate devices is 1 (Fig. 1 (c)). At hop 6 and 9 the degree distribution has two peaks with the second considerably higher than the first. In Fig. 2 (b), the decay of the curve in the second eigenvalue is more evident and the

subsequent eigenvalues are small. In general, we can notice that the first 10 eigenvalues contain most of the information on the routing matrix.

In this preliminary analysis we have discussed on the structures of the connectivity tree of the vantage point. The properties of most of the paths can be estimated from a smaller number of monitored paths by using the shared links. This observation applies very well to the impact factor where a link failure impacts all paths passing through it. The drawback is that the vantage point should construct the routing matrix, but this is often unfeasible since the connectivity tree can change due to routing updates and this is in any case not scalable. Thus, we first propose an estimator for the impact factor that not only reduces the number of monitored paths, but does not require any topological information. We leverage the monitoring and detection capabilities of the vantage point to infer the seriousness of the anomaly and the quality of access. This is summarized and quantified by the impact factor which ranges between 0 and 1, with 0 being the normal case and 1 the case when all destinations are impacted by the anomaly. The quality of the Internet access can be modeled as one minus the impact factor.

### III. IMPACT FACTOR ESTIMATOR

In this section, we define an unbiased estimator for the impact factor and we formalize the problem subsequently addressed. We consider a single end user, which is the vantage point for monitoring the local status of the network, and we model the seriousness of an anomaly observed at this point. For this analysis, we do not differentiate among different network anomalies occurring at the same time but we consider a general case when there might be more than one anomalous link under the general term anomaly. We also do not differentiate between a congested or unavailable link since both problems cause performance degradation of the end users' traffic; we use them equally to refer a network anomaly.

Let  $\mathcal{D}$  be the set of all possible destinations in the network,  $n_D = |\mathcal{D}|$  being their number. Let  $\mathcal{L}$  be a set of  $n_L = |\mathcal{L}|$  landmarks randomly chosen among the destinations evenly distributed ( $\mathcal{L} \subseteq \mathcal{D}$ ), whose paths are continuously monitored by the vantage point to detect network anomalies. Specifically, the vantage point sends probing packets to observe and actively measure the state of the links of the paths to the landmarks so that it can determine whether they are still accessible with an acceptable service level, or normal quality of the connection. For instance, the vantage point can collect the round-trip times

(RTTs) to the landmarks and consider that a shift in the delay means deviation from the normal conditions of the path. As a general rule, the connection to a destination is not normal if this destination is reached via one or more anomalous links.

We now formalize the problem as follows. Can we define an unbiased estimator of the impact factor leveraging the observations made on the path to the landmarks without knowing any information about the topology? We assume that an anomalous event causes the traffic to a destination to experience a high delay, or a destination to be unreachable. Let  $n_u \leq n_D$  denote the number of all such destinations, then we define the *impact factor* as  $I_f = \frac{n_u}{n_D}$ . Now we want to determine an unbiased estimator  $\hat{I}_f$  of the impact factor computed over  $n_L \leq n_D$  observations, given that a smaller  $n_L$  reduces the overhead of the measurements. Based on the analysis presented in Section II, the estimator should account for an approximation of the connectivity tree to estimate the fraction of destinations interested by the network anomaly.

Let's consider an anomalous event. We construct the vector  $X = \{X_1, X_2, \dots, X_{n_L}\} \in \{0, 1\}^{1 \times n_L}$ .  $X_l$  indicates the result of the observation made by probing landmark  $L_l$ :  $X_l = 1$  if the probing packets to landmark  $L_l$  traverse the anomalous link and  $X_l = 0$  otherwise. By definition, the probability that the vantage point probes the path containing one or more anomalous links is the impact factor  $I_f$  itself. Now, we define the random variable  $S_L = \sum_l X_l$  that indicates the total number of paths to the landmarks that traverse the anomalous link. Clearly,  $S_L$  follows a Binomial Distribution  $B(n_L, I_f)$ : the probability of probing the anomalous link is the impact factor, the landmarks are randomly selected, and the measurements on the paths to the landmarks are independent  $n_L$  trials where  $S_L$  indicates the number of "successes". By the law of large numbers, the impact factor can be defined as

$$I_f = \lim_{n_L \rightarrow \infty} \frac{S_L}{n_L}, \quad (1)$$

since  $n_L$  is the number of landmarks picked randomly and we are modeling the problem with the Binomial distribution where the landmarks are randomly sampled with replacement. Note that at this point we do not make any assumption on the selection of the landmarks and the same path can be selected more than once in constructing the vector  $X$ . The expected value and the variance of the binomial random variable  $S_L$  are  $E[S_L] = n_L \cdot I_f$  and  $VAR[S_L] = n_L \cdot I_f(1 - I_f)$  respectively. It follows that an unbiased estimator of  $I_f$  is

$$\hat{I}_f = \frac{S_L}{n_L}. \quad (2)$$

It is very easy to prove that this estimator is the maximum likelihood estimator for the impact factor given the observed vector  $X$  whose sum of elements is  $S_L$ . This estimator has variance equal to  $\frac{I_f(1-I_f)}{n_L}$ , which drops to zero when the number of landmarks increases to large values.

#### A. Estimation of the number of landmarks

We now study the minimum set of landmarks satisfying a given accuracy of the estimator. Let's consider two small

quantities  $\epsilon$  and  $\alpha$  characterizing the confidence interval and the significance level of the estimator respectively. Now we want to determine the minimum number of landmarks to be monitored so that the  $\hat{I}_f = \frac{S_L}{n_L}$  can be used as an estimator for the impact factor  $I_f$  in a confidence interval equal to  $(\frac{S_L}{n_L} \pm \epsilon)$  with a significance level  $100(1 - \alpha)\%$ . For large  $n_L$ , e.g.  $n_L > 30$ , thanks to the Central Limit Theorem the Normal distribution  $N(n_L \cdot I_f, n_L \cdot I_f(1 - I_f))$  can approximate the Binomial distribution  $S_L$ . For a small value of  $n_L$  we refer to the t-student distribution being aware of the error of this approximation. At this point we are interested in computing the order of magnitude of the number of landmarks and we will investigate this problem in more details in Section V-B where we analyze the empirical error of the approximation by varying  $n_L$ . Thus, the confidence interval of the estimator becomes  $(\hat{I}_f \pm z_{1-\alpha} \sqrt{VAR[\hat{I}_f]})$ , where  $z_{1-\alpha}$  is the  $100(1 - \alpha)$ -th percentile of the standard normal distribution and  $VAR[\hat{I}_f] = \frac{VAR[S_L]}{n_L^2}$ . By setting this interval to the target accuracy  $\epsilon$  and substituting  $VAR[S_L]$  by its expression, one can easily obtain that the number of landmarks should satisfy  $n_L \geq (\frac{z_{1-\alpha}}{\epsilon})^2 \cdot I_f \cdot (1 - I_f)$ .

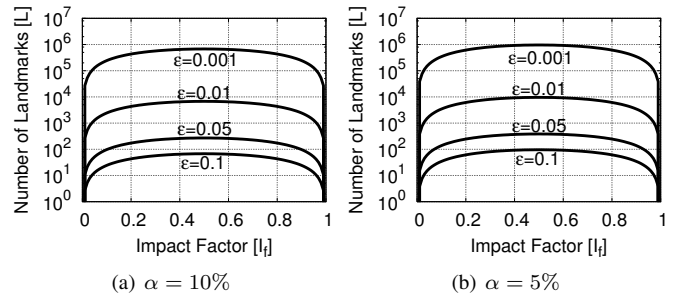


Fig. 3. Minimum number of landmarks on a logarithmic scale to reach a significant level  $100(1 - \alpha)\%$  for the estimator.

Fig. 3 plots the minimum number of landmarks to be randomly selected as a function of the impact factor for different significance levels such that an estimated impact factor lies in a confidence interval defined by  $\epsilon$ . We observe that the curves are symmetric around  $I_f = 0.5$  and for an impact factor close to 0, or 1, the minimum number of landmarks needed can be very small. In particular, the figures show that fewer landmarks are needed for a reliable estimation if the anomaly affects few destinations or a large set of them, in comparison to the intermediate case when the entropy of the binomial distribution is maximal. In practice, if there is no anomaly in the network, all landmarks report normal conditions, hence few of them are enough. In the opposite case if the anomaly is severe, all network destinations will be unreachable; in these cases even 1 landmark is enough to monitor the status of this access link. For values close to 0 or 1, a small confidence interval for a given significance level is sensitive to both the small deviation of the impact factor and to the number of landmarks, leading to the sharp change of the curves. We can summarize the results and say that few landmarks are sufficient to assess the quality of its access link (really good or really bad), this quality being one minus the impact factor we define.

## IV. METHODOLOGY

The validation of our model is performed on a simulator built on real data used to construct the local topology for each vantage point, i.e., the connectivity tree, so that we are able to control the experiments and study the impact factor in details.

### A. Experimental data

In our analysis we use the data set of the iPlane project [12] gathered on August 22, 2009. iPlane collects traceroute measurements from 199 vantage points, located mostly on Planet-Lab [15], to  $\sim 142,000$  public IP addresses with more than one destination in the same IP prefix. We resolve the aliases and we only keep one destination in each  $/24$  network prefix. From this set, we filter out vantage points that have an incomplete set of measurements or few destinations; although these trees are also interesting to analyze, we prefer to avoid possible bias in the evaluation of the estimator. We also filter out the unreachable destinations and the traceroutes that exhibit some known problems: routing loops, destination address in the middle of the path, hidden and unknown intermediate routers. The final data set for the experiments, unless specified, consists of 137 vantage points and on average 29,000 paths to unique destinations and 75,000 unique links per vantage point.

In this paper, we make no explicit assertions about the extent to which this data set represents the actual core of the Internet since there is no authoritative source for that information. While the limitations of using Planetlab for investigations of the Internet topology are known [16], we argue that both the scope and content of the data are appropriate for our analysis since the goal of this work is primarily methodological in nature. Furthermore, the fact that our data are gathered during a single day is not a limit of the analysis as our objective is to use a realistic router-level topology as seen by an end user.

### B. Anomaly simulator

Our goal is to analyze network anomalies in a real topology to determine the real impact factor and validate the estimator defined in Section III. We have deployed a simulator that constructs the connectivity tree from the traceroute measurements and generates some controlled anomalies. This choice is determined by the fact that network anomalies in reality might not happen at regular intervals to be able to study them by real experimentations. Experimenting directly on the real Internet does not allow a complete study of the impact factor since network anomalies are hard to reproduce; further, it is infeasible to compute in reality the impact factor to assess the goodness of the estimator. We will discuss more about it in Section VI. In summary, the simulator allows: (i) to control directly the anomaly; (ii) to perform a sensitivity analysis on the number of landmarks; (iii) to determine the uncertainty in localizing the network anomaly; (iv) to understand the limitations and advantages of the estimator in detail [17]. Nevertheless, we are interested in evaluating real scenarios, thus, we define our simulator on top of real data to understand how the impact factor estimator performs in reality.

The simulator processes the data of each vantage point separately for constructing the connectivity tree. We fix the number of landmarks used for the experiment and each vantage point randomly selects the set among the destinations. This set is the same for the entire duration to guarantee that the experiments are as close as possible to the real case. In fact, we assume that a vantage point selects its landmarks and sticks with this choice to profit from temporal correlation for detecting anomalies on these paths and for estimating the Internet access quality on the time series of the impact factor.

We generate anomalies by eliminating links from the connectivity tree to simulate broken or heavily congested links and at fixed instants of time the vantage point constructs the vector of observations, i.e.,  $X$ , based on the results of the detection mechanism: 1 in case of detected network anomaly and 0 otherwise. For each anomalous event and for each vantage point, we check the tree to compute the real and the estimated impact factor as the fraction of destinations and landmarks interested by the anomaly, respectively. By considering that the destinations in the iPlane data are not biased against some specific ASes, this calculated impact factor can be safely assumed to be close to the real one that the vantage point would observe in the wild.

The major difficulty in our study comes from the uncertainty in the set of feasible impact factor values in reality. This uncertainty adds to the error introduced by the choice of a limited set of landmarks to probe. The location of the anomaly by inverting landmark-based probing is another interesting question to handle. Without loss of generality, and supposing all links can be anomalous with the same probability, the present analysis should shed the light on how end-to-end probes can evaluate the quality of an Internet access.

## V. IMPACT FACTOR ANALYSIS

In this section, we study the impact factor on real traces and determine how the impact factor varies. We experimentally analyze the minimum number of landmarks to be monitored for estimating the impact factor given a satisfactory accuracy.

### A. Real Impact Factor

We start by analyzing the real impact factor considering anomalies caused by one anomalous link at a time and iterating for all links in the connectivity trees. The vantage points select randomly the set of landmarks and use the same set for all the measurements. In accordance with the way we define our estimator, no consideration on the best or worst set of landmarks is made.

In Section II, we analyze the connectivity graph and show that the distribution of the links is function of the depth of the connectivity tree. Let's first consider a balanced tree, then the impact factor is inversely proportional to the hop distance of the anomalous link since the farther is the anomaly, fewer destinations are connected through the anomalous link. For instance if the connectivity tree is balanced and nodes have a degree  $k$ , then we would expect that  $I_f = 1/k^h$ , where  $h$  is the number of the hops. In particular, the impact factor is  $1/k$  at

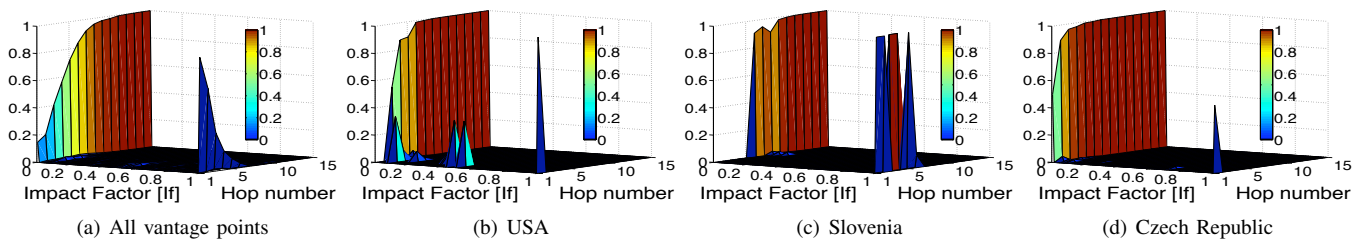


Fig. 4. Probability density function per hop of the Impact factor on real data

the first hop with probability  $1/k$  and 0 with probability  $(k - 1)/k$ . Thus, the presence of intermediate values between 0 and 1 is determined by the degree of the routers in the connectivity tree and by the links at few hop distance that connect only few destinations. Fig. 4 plots the probability density function of the impact factor for different values of the hop distance of the anomalous link (the number of hops reaches 30 but the plots are truncated for clearness). The plots show that the intermediate values of the impact factor are possible only for close anomalies.

Fig. 4 (a) plots the probability density function of the average impact factor of all 137 vantage points used for the experiments when the anomaly is located at a specific  $n$ -hop distance from the vantage point. We notice that averaging among all possible vantage points compensate for connectivity trees that have a high or small degree for the intermediate routers (see Fig. 1). Fig. 4 (a) shows that at the first hop the impact factor is 1 with 0.85 probability and low probability is associated to other possible values of the impact factor. This means that the connectivity tree constructed by averaging all vantage points is not balanced.

Fig. 4 (b), (c), and (d) plot the probability density of the impact factor for the vantage points located in USA, Slovenia, and Czech Republic respectively, whose connectivity trees have been studied in Section II. The probability density of the impact factor for the USA vantage point shows that the impact factor at the first hop is 1 and anomalies at the second hop span almost all possible values below 0.5. This confirms the analysis of Section II about the degree distribution, Fig. 1 (b), and the eigenspectrum of the connectivity tree, Fig. 2 (b), since the tree is almost balanced after the first hop. Fig. 4 (d) shows the results for the vantage point located in Czech Republic. In this case, the impact factor has equal probability to be close to 1 or 0; this is because the vantage point has less shared links in the first hops but with high degree routers, thus its impact factor becomes almost 0 at the second hop. An opposite behavior is however shown in Fig. 4 (c). The vantage point located in Slovenia has a degree distribution of 1 for the first 5 hops and a peak at hop 6 (see Fig. 1 (c)). It follows that an anomaly located at these 5 hops impacts seriously the end user's traffic, but the ones at hop 6 and more have almost negligible impact.

## B. Number of landmarks

In Section III we probabilistically compute the minimum number of landmarks considering the range of all possible values of the impact factor. However, we have noticed that

the impact factor presents values with a non uniform distribution in reality, in particular the distribution has the mass concentrated on low or high values.

We use the impact factor distribution of Fig. 4 (a) to calculate the minimum number of landmarks sufficient to estimate the impact factor itself for anomalous links at a specific. We plot the cumulative distribution function in Fig. 5. We analyze the significance level for a confidence interval defined by  $\epsilon = 0.01$ , in plots (a) and (b), and the confidence interval for a 0.95 significance level, in plots (c) and (d); the bottom of the plots shows the colored cumulative distribution. In general, we observe that the minimum number of landmarks is more sensitive to the size of the confidence interval, plot (d). Fig. 5 (b) shows that for a 99% significance level, 75 landmarks can estimate the impact factor considering an error of 1% for an anomalous link at any hop distance.

It is worth noticing the sharp increase of the number of sufficient landmarks. The number at which we have the increase is different and it reflects the distribution of the impact factor. Indeed, a far anomalous link has an impact factor close to 0 most of the time such that the required number of landmarks is small; similarly an anomaly at the first hop has impact factor 1 and few landmarks are enough also in this case. On the contrary, when the impact factor is between 0 and 1, there is higher uncertainty since the estimator is more sensitive to the location of the landmarks and the links crossed, thus a larger number of landmarks is required for the estimation. This is in line with the theoretical results in Fig. 3 of Section III-A.

Considering the cost of continuously monitoring the landmarks and the most probable values of the impact factor, we can conclude that in reality the end user does only need few landmarks and 10 (Fig. 5 (a)) is a good compromise to have an indication of the quality of the access network. In the rest of this section we analyze the estimated impact factor in two particular cases: 10 and 100 landmarks.

## C. Estimated Impact Factor

The plots in Fig. 6 show the probability density function of the estimated impact factor when 10 and 100 landmarks are used; these are the plots (a)-(d) and (e)-(h), respectively. Fig. 4 (a) and (b) show that the estimator is able to infer correctly the impact factor for all cases with only 10 landmarks. 10 landmarks normally would give a high error in the estimation for an intermediate value of the impact factor. However, in reality the most probable values of the impact factor are close to 0 and 1, hence, a small number of landmarks is in practice sufficient (see Fig. 5). This is evident for the vantage point

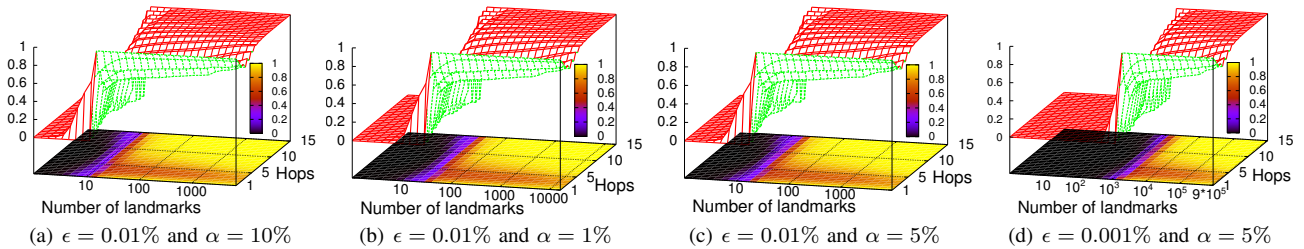


Fig. 5. Cumulative distribution function of the minimum number of landmarks based on the real impact factor.

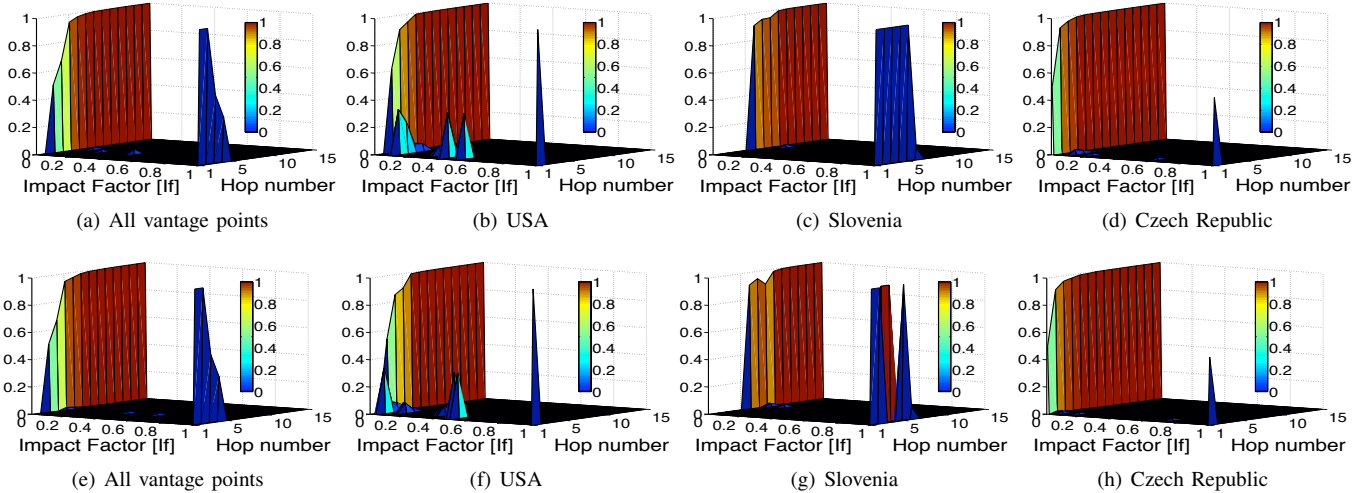


Fig. 6. Probability density function of the estimated impact factor with 10 and 100 landmarks in plots (a-d) and (e-f), respectively.

located in Slovenia and Czech Republic (Fig. 6 (c) and (d)) since the real impact factor has few values between 0 and 1. The vantage point of Fig. 4 (b) has its traffic equally divided between the links at the first hops causing the estimator to suffer from a poor selection of the landmarks. In reality, Fig. 4 (b) shows that 10 landmarks give a good estimate of the impact factor. The estimator becomes more accurate when the number of landmarks increases. Fig. 6 (e-h) show that an estimator with 100 landmarks models well the impact factor in all cases.

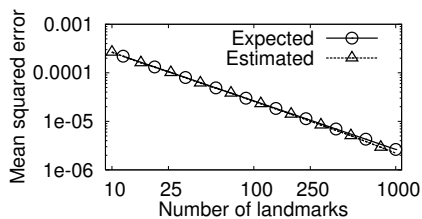


Fig. 7. Mean square error of the estimated impact factor in log-log scale

Fig. 7 shows the mean square error of the estimator versus the number of landmarks for a connectivity tree up to hop 7, since these are the hops where there is more uncertainty for the value of the impact factor in case of an anomaly. We compute the results for 1,000 different selections of landmarks, over all vantage points and anomalies. Fig. 7 plots the mean of the estimated and expected square error, i.e.,  $\frac{I_f \cdot (1 - I_f)}{n_L \cdot n}$ , where  $n$  is the number of different anomalies we consider. The curves almost overlap and the mean square error is of the order of 0.1% for 10 landmarks; further, we have computed from the data a maximum error of 1% in 99% of the cases.

We can summarize the results of this section saying that

an end user can estimate the quality of Internet access with few landmarks. We have shown that 10 randomly selected landmarks estimate in most of the cases the impact factor and this depends on the connectivity tree of the vantage point. Nevertheless, 10 is a sufficient number to have an idea of how the anomalies impact the traffic. An important finding is that an anomaly affects the Internet access of an end user if anomalies are located at the first hops, otherwise the impact factor is almost zero. This validates the use of the impact factor as a metric to distinguish between an external problem to the ISP causing the malfunctioning of the Internet access or an internal problem of which the ISP should be responsible for.

## VI. DISCUSSION

Critical to the real implementation of our model is the anomaly detection phase, since techniques are prone to errors in the measurements due to the intrusiveness of probes, ISPs' behaviors and uncooperative nodes, such as routers which block probing packets. In our analysis we have assumed that the vantage points detect network anomalies on the monitored paths without error. Whether this is desirable, in practice the detection phase is not accurate and the filters might not detect network anomalies or generate false alarms. Hence, the measurement error affects the impact factor estimator. We can refer to the performance of the detection filters in terms of sensitivity (or true positive rate TP), i.e., the actual network anomalies correctly reported, and specificity (or true negative rate TN), i.e., normal path conditions correctly reported. Then, the corrected impact factor estimator is simply  $cI_f = I_f \cdot TP + (1 - I_f)(1 - TN)$ . In practice, this means

that a larger number of landmarks is required to compensate the detection error such that the sample size becomes  $n_L \geq \left(\frac{z_{1-\alpha}}{\epsilon(TN+TP-1)}\right)^2 \cdot cI_f \cdot (1-cI_f)$  to achieve a confidence interval defined by  $\epsilon$  with a confidence level  $100(1-\alpha)\%$  [18]. In the future we plan to investigate in reality the use of detection filters. We also plan to study if it is possible to profit from the impact factor distribution to improve the accuracy of the estimator or if information about previous impact factors can be used to compensate for the misclassification error while maintaining the number of landmarks constant.

Another aspect worth of consideration for the real implementation of our framework is the definition of landmarks. The performance of the estimator also depends on the position of the landmarks and a bad selection can significantly reduce the accuracy. For instance, a random selection of the landmarks does not consider the users' traffic preferences and the estimator is unbiased with respect to a random destination. On the contrary, if the traffic of the user is predominant toward some IP networks connected by few links, monitoring landmarks in only these autonomous system does not help in inferring the impact factor as most of the destinations are left out. We plan to use passive measurements to bias the selection of landmarks toward some destination networks more interesting for the end user.

Most of the detection tools probe continuously the paths so that the landmarks should be willing to accept this extra traffic. Our belief is that end users are willing to collaborate if this serves to troubleshoot anomalies and to quantify the quality of Internet access. For example, Grenouille [1] already provides free software to measure a set of metrics and the end users participate actively to monitor their access network. We also envision the presence of multiple collaborative clients, but we also expect the ISPs to provide some dedicated network devices close to the access networks to give a guarantee of their services. The collaboration of different end users is also subject of future work. In this paper we have studied each vantage point separately and in the future we plan to investigate how the collaboration of the users and the spatial correlation of the end-to-end measurements can improve the accuracy of the estimator and the localization of the anomaly.

We also plan to evaluate the estimator on data sets that resemble better the connectivity of the users, such as the DIMES [19]. The iPlane data contains mostly vantage points located on PlanetLab which are connected to the Internet through non-commercial autonomous systems. In this study we have highlighted the characteristic of the connectivity tree of the vantage points in such a way that we could detach the properties of the estimator from the structure of the topology. We notice that the real values of the impact factor are linked with the connectivity of the end user; thus, we want to understand whether we could exploit specific properties to enhance the accuracy of the estimator.

## VII. CONCLUSIONS

In this paper we propose a novel metric for the evaluation of the Internet access to estimate whether the ISP is fulfilling the

Service Level Agreement. In contrast to previous work we focus on estimating the real impact of network anomalies on the user Internet access and we define the impact of an anomaly as being the fraction of destinations and servers unreachable across the Internet. We have defined a probabilistic model to infer the impact factor from a set of end-to-end measurements to random destinations, substantially smaller than that needed for exact monitoring, without assuming the knowledge of the network topology. Our work is methodological in nature and exploits the redundancy in links of the paths from the source to the destinations. We have used a set of simulations based on real traces to evaluate the estimator in practice and the results confirm the validity of the impact factor as a metric, showing that 10 landmarks are already sufficient to provide an indication of the quality of access. Often, the estimator has a small error which depends mostly on the size of the landmarks set. The results of this paper shed light on other practical issues and opens new research opportunities. From the point of view of the probabilistic model, we have defined the estimator without considering the temporal correlation of the measurements, which can be exploited to reduce the number of landmarks while maintaining the same accuracy.

## REFERENCES

- [1] "<http://grenouille.com/>."
- [2] P. Barford, N. Duffield, A. Ron, and J. Sommers, "Network performance anomaly detection and localization," in *IEEE INFOCOM*, Apr 2009.
- [3] A. Lakhina, M. Crovella, and C. Diot, "Characterization of network-wide anomalies in traffic flows," in *ACM IMC*, 2004.
- [4] —, "Diagnosing network-wide traffic anomalies," in *ACM SIGCOMM*, 2004.
- [5] —, "Mining anomalies using traffic feature distributions," in *ACM SIGCOMM*, 2005.
- [6] Y. Mao, H. Jamjoom, S. Tao, and J. M. Smith, "Networkmd: topology inference and failure diagnosis in the last mile," in *ACM IMC*, 2007.
- [7] R. Castro, M. Coates, G. Liang, R. D. Nowak, and B. Yu, "Network tomography: Recent developments," *Statistical Science*, vol. 19, no. 3, pp. 499–517, 2004.
- [8] F. Lo Presti, N. G. Duffield, J. Horowitz, and D. Towsley, "Multicast-based inference of network-internal delay distributions," *IEEE/ACM Trans. Netw.*, vol. 10, no. 6, pp. 761–775, 2002.
- [9] A. Dhamdhere, R. Teixeira, C. Dovrolis, and C. Diot, "Netdiagnoser: troubleshooting network unreachabilities using end-to-end probes and routing data," in *ACM CoNEXT conference*, USA, 2007, pp. 1–12.
- [10] M. Costa, M. Castro, A. Rowstron, and P. Key, "Pic: Practical internet coordinates for distance estimation," in *IEEE ICDCS*, Japan, 2004.
- [11] D. B. Chua, E. D. Kolaczyk, and M. Crovella, "Efficient monitoring of end-to-end network properties," in *IEEE INFOCOM*, March 2005.
- [12] "<http://iplane.cs.washington.edu/>."
- [13] C. Gkantsidis, M. Mihail, and E. W. Zegura, "Spectral analysis of internet topologies," in *IEEE INFOCOM*, 2003.
- [14] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the internet topology," in *SIGCOMM*, 1999.
- [15] "<http://www.planet-lab.org/>."
- [16] N. Spring, L. Peterson, A. Bavier, and V. Pai, "Using planetlab for network research: myths, realities, and best practices," *SIGOPS Operating System Review*, vol. 40, no. 1, pp. 17–24, 2006.
- [17] H. Ringberg, M. Roughan, and J. Rexford, "The need for simulation in evaluating anomaly detectors," *SIGCOMM Computer Communication Review*, vol. 38, no. 1, Jan 2008.
- [18] E. Rahme and L. Joseph, "Estimating the prevalence of a rare disease: adjusted maximum likelihood," *The Statistician*, vol. 47, no. 1, 1998.
- [19] Y. Shavitt and E. Shir, "Dimes: let the internet measure itself," *SIGCOMM Comput. Commun. Rev.*, vol. 35, no. 5, pp. 71–74, 2005.