# On the end-to-end connectivity evolution of the Internet

**Thiago B. Cardozo[1], Ana Paula C. Silva[1], Alex B. Vieira[1], Artur Ziviani[2]**

[1]Departamento de Cincia da Computao – Universidade Federal de Juiz de Fora (UFJF)
Juiz de Fora – MG – Brasil

[2]Laboratrio Nacional de Computao Cientfica
Petrpolis – RJ – Brasil

thiago.boubee@ice.ufjf.br

{anapaula.silva,alex.borges}@ufjf.edu.br,ziviani@lncc.br

***Abstract.*** *The Internet is a system under continuous evolution. In this paper, we characterize and analyze the recent end-to-end connectivity evolution of the Internet, comparing key end-to-end performance metrics from two distinct periods separated by five years. Our findings show that the average path length distribution slightly changed from 2006 to 2011, but the delay distribution actually became worse, with a* $45\%$ *increase in path delay from 2006 to 2011. This directly affects network performance and degrades user experience. Furthermore, we show that path diversity decreased, and accordingly, distinct paths became slightly more similar. This result has a direct impact on routing algorithms that try to explore path diversity to become more fault-tolerant.*

## 1. Introduction

The Internet currently plays a key role in our communication and collaboration infrastructure. Such a system is under continuous evolution over the years, with many nodes and links being created while many others are suppressed.

Moreover, the network performance of the Internet is critical to communications and on-line services [Lee et al. 2010]. From home entertainment to business systems, Internet metrics, such as delay and hop count, impacts user satisfaction and companies profit. Therefore, in order to develop better network models and systems, it is important to understand Internet behavior and analyze trends, in particular how relevant metrics change over time.

In this paper, we study the end-to-end connectivity evolution of the Internet over recent years. This allows a fine-grained characterization, in contrast to most previous AS-level studies [Dhamdhere and Dovrolis 2011, Siganos et al. 2002]. We compare key performance metrics from two distinct periods separated by five years. We consider data from 2006 and 2011, including the end-to-end path latencies, path number of hops, and path geographic dispersion. We further propose a method based on string edit distance to characterize the path diversity and measure how different paths are in both periods.

There are several studies on Internet topology measurement and modeling [Haddadi et al. 2008, Oliveira et al. 2007, Zhou 2006]. Most of them rely on Internet probes for a limited time period and view the Internet topology from the AS perspective, ignoring the end-to-end relationship. Moreover, most of the earlier work that claims to characterize and model Internet dynamics focuses on the evolution of topological properties, such as degree distribution, clustering coefficient, or graph diameter, disregarding user-centric metrics, such as end-to-end latency or path size. These are very important metrics we consider in this study because they impact user satisfaction and possible interdomain traffic costs.

Our results show that key aspects of the Internet are actually getting worse in recent years. The average path length distribution remains almost identical from 2006 to 2011, but the delay distribution clearly became worse. In this case, we observe a $45\%$ increase in path delay from 2006 to 2011. This directly affects network performance and degrades user experience. Furthermore, we show that path diversity decreased, and accordingly, distinct paths became slightly more similar. Two distinct paths to the same end points in 2006 used to be $63\%$ different, but in 2011, this difference falls to $54\%$. This result has a direct impact on routing algorithms that try to explore path diversity to become more fault-tolerant. Overall, our study helps understanding the recent evolution of the Internet connectivity and the consequent user experience.

The remainder of this paper is organized as follows. Section 2 introduces the datasets we use as well as the adopted sampling methodology. In Section 3, we present the considered evaluation metrics and the obtained results. Section 4 briefly reviews related work. Finally, Section 5 concludes the paper and discusses future work.

## 2. Sampling Internet Paths

### 2.1. Datasets

We rely our work on a set of Internet's historical data from the CAIDA project[1]. The data CAIDA provides have been available for over a decade. This data is a traceroute like output from where we are able to gather end-to-end Internet paths and delay measurements. We use data from a set of North America and Europe monitors as most of 2006 monitors were located in these continents.

The CAIDA datasets obviously do not provide a complete Internet map. Nevertheless, they still provide a suitable standpoint to base our investigation with a representative delay distribution of the Internet. Moreover, these data sets are among the largest data archives publicly available and are constantly updated.

We acquired all available network traces in the CAIDA project taken in all months of 2006 (Skitter project) and 2011 (Ark project). In such traces, there are 19 and 39 monitors labeled as a source in 2006 and 2011, respectively. Out of these, 13 and 18 sources are in North America and 6 and 21 are in Europe in 2006 and 2011, respectively. Each year contains an expressive number of traceroute outputs. For instance, we have about $800$ million outputs for 2006 and more than $1$ billion outputs for 2011.

### 2.2. Sampling Methodology

We use the datasets in two ways. First, we process a subset of available data from CAIDA traceroute like output to calculate metrics that can be directly inferred from network probes. For instance, we characterize the number of hops a path has and the path delay. Such metrics are further discussed in Section 3.1. Second, we rebuild a partial view of the Internet reconstructing a graph from a subset of traces we collected from CAIDA. This enables the reconstruction of end-to-end paths between two endpoints and also to infer new end-to-end paths not present in the original set of CAIDA outputs. In this case, we can analyze the path diversity between two endpoints.

We randomly choose $n$ unique traceroutes we have available in the original CAIDA outputs. We expect that these $n$ unique end-to-end measures qualitatively represent a much larger set because hosts in the same /24 block are likely to experience similar performance, such as network delays and packet losses. In this work, we choose $n =$ 200 million as the sample size for improving accuracy of our end-to-end connectivity analysis. Lee et al [Lee et al. 2010] show that a sample size of $n =$ 50,000 $\sim$ 60,000 presents very small errors in median estimation for a very tight (99%) confidence interval.

We represent our partial view of the Internet as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of nodes we found in all traces we select and $\mathcal{E}$ is the set of links between a pair of nodes. We reference $\mathcal{G}_{2006}$ and $\mathcal{G}_{2011}$ as the graphs modeling the Internet topology from years 2006 and 2011, respectively. The Internet graphs we rebuild from $m = 100,000$ randomly traceroutes available in the original CAIDA outputs: $\mathcal{G}_{2006}$ graph has 172,926 nodes and 459,498 links; and $\mathcal{G}_{2011}$ graph has 162,119 nodes and 448,614 links.

In most cases, we rely our characterization in the properties we found in the $K$-shortest paths from the partial view of the Internet. The importance of this analysis is

[1] www.caida.org/

related to the fact that most common routing protocols tend to follow the shortest paths between two nodes on the network (e.g. OSPF). Longer paths, despite their existence, are not used in practice. We randomly select 10 origins and 10,000 destinations to generate the $K$-shortest paths we use in our characterization.

The K-shortest paths problem consists on the determination of a set $\{p_1, ..., p_k\}$ of paths between a given pair of nodes when the objective function of the shortest path problem is considered and in such a way that $c\,(p_k) \leq c\,(p)$ for any $p \in P - P\,(k-1)$, where $P\,(k) = \{p_1, ..., p_k\}$ and $P\,(0)$ is the empty set. The shortest path is the first to be determined, then the second shortest is the second to be determined and so on.

## 3. Connectivity Evolution

In this section, we review the metrics we use to do our characterization (Section 3.1). We also report the results we found during an analysis of a five year interval of the end-to-end Internet connectivity (Section 3.2).

### 3.1. Metrics

Our study relies on the most common metrics used to characterize end-to-end connectivity. For instance, we can describe an Internet path in terms of its size (hops count) and latency (path delay). We also characterize the Internet path diversity. In what follows, we formalize the set of metrics selected for our analysis of Internet connectivity evolution:

- *Path Number of Hops* (or hop count) is the distance between two hosts. We gather the path number of hops from $n = 200$ million CAIDA traceroute outputs and from our partial view, rebuild with $m = 100,000$ CAIDA traceroute outputs. We also analyze the number of hops from the K-shortest path we sample from the Internet partial view.

- *Path Delay* is the round trip time (RTT) between two given hosts. A large path delay may indicate a long path (large hop count) or a link congestion. We analyze the path delay using the RTT we get from $n = 200$ million CAIDA traceroute outputs. We also use the path delay inferred from the partial view ($m = 100,000$ CAIDA traceroute outputs) of the Internet to correlate path properties.

- *Path Diversity* is a metric that reflects the number of alternative routes available between any pair origin-destination. Path Diversity allows quantifying the connectivity richness in the Internet topology. We capture the path diversity characteristics analyzing the differences between the $K$-shortest paths, using graphs $\mathcal{G}_{2006}$ and $\mathcal{G}_{2011}$.

- *Paths Difference*: we quantify the difference between two paths that links two given end points using the minimum *Levenshtein distance*. In other words, the paths difference is the minimum number of operations we have to do in order to transform a path into another one.

- *Path Geographic Dispersion* $D_g$ of a path between two endpoints is defined as the ratio between the total path geodesic distance $D_p$, considering the sequence of different intermediate points that compose an end-to-end path, and the direct real geodesic distance $D_r$ between the two end points of the path, i.e.

$$D_g = \tfrac{|D_p - D_r|}{D_r}.$$

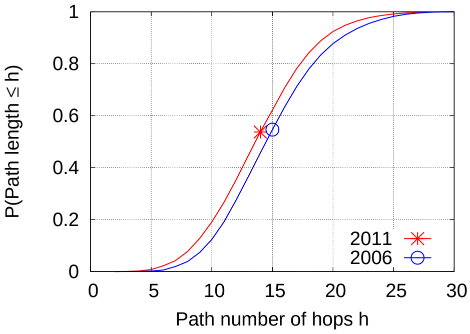


**Figure 1. Path number of hops CDF.**

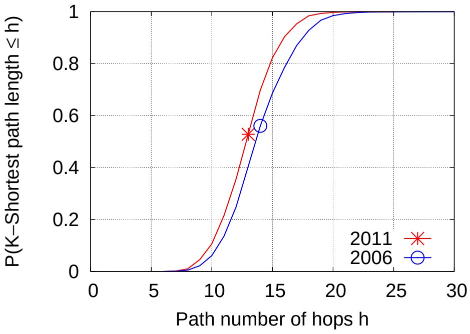


**Figure 2. K=10-shortest path number of hops CDF.**

It is expected that the richer the observed connectivity is, the lesser the geographic dispersion will be, as more direct paths become available. We also use the $K$-shortest paths from the partial view of the Internet (graphs $\mathcal{G}_{2006}$ and $\mathcal{G}_{2011}$) to analyze the path geographic dispersion.

## 3.2. Results

In this section, we analyze the results of our characterization study using the metrics presented in Section 3.1.

### 3.2.1. Path number of hops

We first analyze the path number of hops between two endpoints from our sample of $n = 200$ million CAIDA traces from 2006 and 2011. Interestingly, Figure 1 shows that the cumulative distribution function (CDF) of the paths hops slightly changes from 2006 to 2011, maintaining the same slope of the curves. Despite the new investments in network infrastructure over this 5-years interval, the end-to-end paths have a marginal improvement of one hop. In both cases, 80% of the paths have about 17 hops.

Similar results are reported for AS-level networks [Dhamdhere and Dovrolis 2008]. As the network grows and path lengths remain unchanged, we may think that the network is becoming denser. A denser network may have a richer connectivity, as it should have more paths available between two endpoints.

We also analyze the size of all $K$-shortest-paths from the Internet partial view. Figure 2 shows that 2011 paths are slightly shorter than 2006 paths. Almost 80% of paths have no more than 16 hops and we did not notice any $K$-shortest-path larger than 25 hops in 2011.

In Figure 3, we take a closer look at the individual $K$-shortest-paths. The behavior is the same, and as expected the $K = 1$ path is the shortest and the path length remains stable as $K$ increases to 10, with 2011 paths being slightly shorter than 2006 paths.

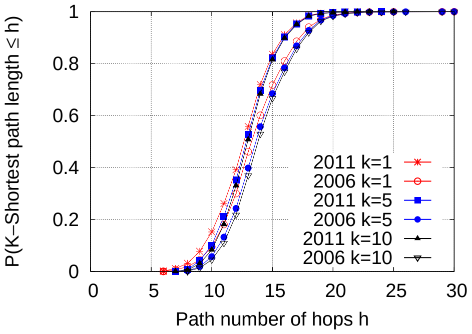


**Figure 3. K-shortest path length CDF.**

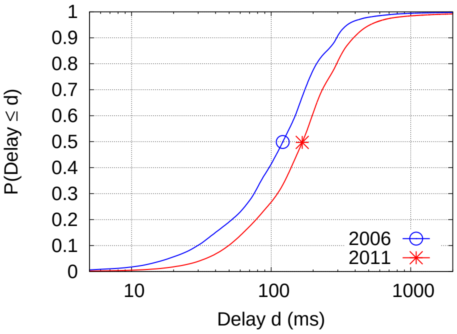


**Figure 4. Path delay CDF.**

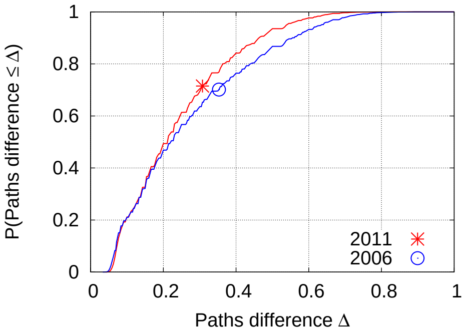


**Figure 5. K=10-shortest paths difference CDF.**

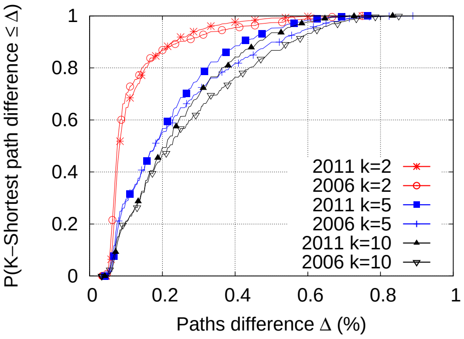


**Figure 6. K-shortest paths difference CDF.**

### 3.2.2. Path Delay

Figure 4 presents the CDF of path delays in log scale. The path delays in the Internet are becoming slightly worse comparing 2011 to 2006. In fact, this shows that in 2006, about $50\%$ (median) of the paths present a delay of at least $122\ ms$, while in 2011 the median delay grew about $37\%$ to almost $168\ ms$. Considering the $95\%$-percentile, the delay increased from about $352\ ms$ in 2006 to about $512\ ms$ in 2011 (a $45\%$ increase in five years).

Our results reinforce that Internet performance is getting worse lately, as reported in [Lee et al. 2010]. A possible reason to this may be a phenomenon known as *bufferbloat*, recently reported in [Gettys and Nichols 2012], where the excess of packet buffering in the network causes high latency and jitter. Furthermore, we also note that the number of virtualized systems is growing. The heavy network usage from virtual machines can introduce delays as high as 100 ms to round-trip times [Whiteaker et al. 2011]. As consequence, the network overall latency also increases.

### 3.2.3. Paths Difference

Figure 5 shows the CDF of the path differences between the $K$-shortest-paths connecting two endpoints. Results suggest the paths difference is diminishing over this 5-year period. In 2011, less than 10% of the paths present a difference larger than 50%, whereas in 2006,

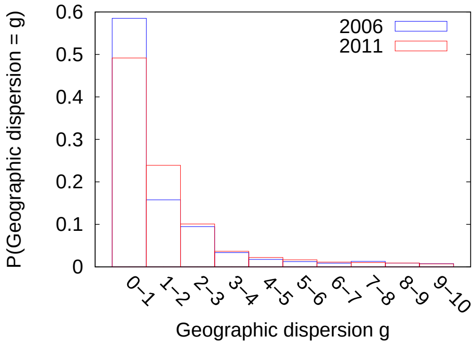


**Figure 7. Probability distribution function of the geographic dispersion of the K-shortest paths.**

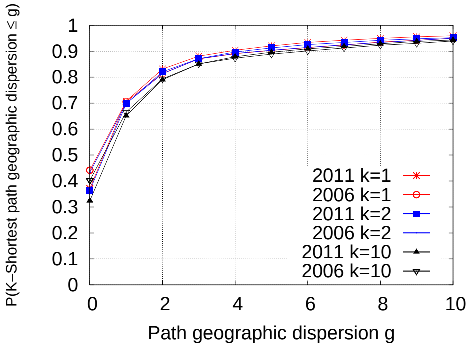


**Figure 8. K-shortest paths geographic dispersion.**

this number reaches 15% of the paths.

A key question that arises from this result is "*if the Internet is getting more dense, why are the shortest paths more similar?*" We believe that the diversity grew around the available shortest paths. Being so, the paths share a lot of nodes in common, having long overlaps along the way, which in turn makes them more similar.

Figure 6 shows that as $K$ increases, the larger is the paths difference. For instance, 80% of the 2011 paths differences are lower than 15% for k=2, whereas the paths differences are larger 36% for k=10. Thus, routing protocols that consider more paths than the shortest one are be able to exploit more diversity and achieve better fault-tolerance.

### 3.2.4. Path Geographic Dispersion

We obtain a host geographic location using a free database provided by MaxMind[2]. We are capable of obtaining a city-level granularity, with host latitude and longitude.

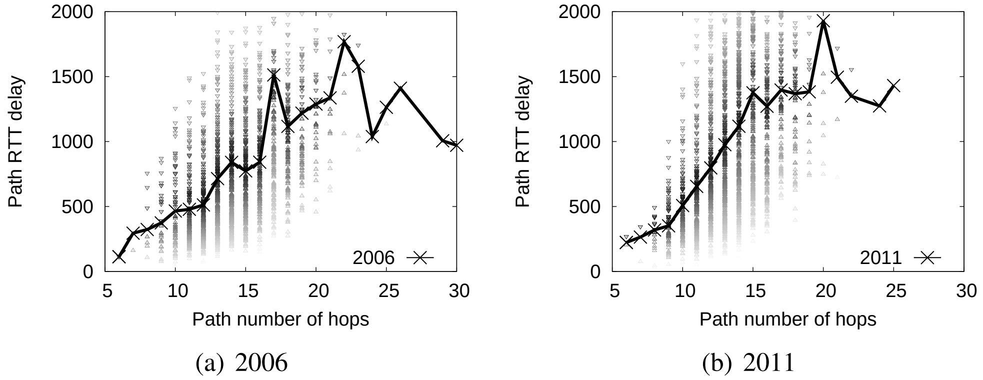
Figure 7 shows that most paths have dispersion from $0$ to $3$, where $0$ is almost a direct geodesic line.

We conjecture that geographic dispersion tends to be relatively small because most links in a path tend to follow a straight line. Moreover, long links that we expect to impact a geographic dispersion (as transoceanic cables), occur in long paths. Thus, a single transoceanic link among all others does not impact the overall dispersion.

Finally, 2011 has about 10% (Fig. 7) of paths with larger dispersion than 2006. This may suggest that small Internet providers are changing their local links to long distance links. Most large network operators may typically divert their traffic to further distances due to economical reasons.

We also show the CDF of geographic dispersion for each $Kth$ path in Figure 8. We note that, as $K$ increases to $10$, the geographic dispersion remains stable. In this case, almost 90% of paths have dispersion lower than $4$.

[2] `http://www.maxmind.com/`



(a) 2006

(b) 2011

**Figure 9. Correlation between path delay and path number of hops.**

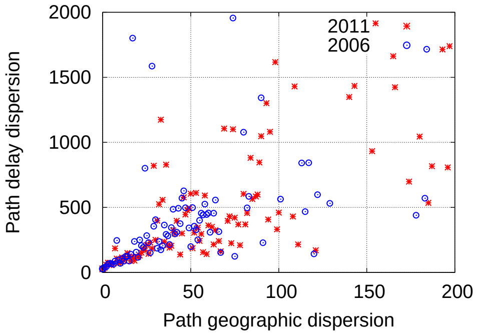


**Figure 10. Correlation between path delay dispersion and geographic dispersion.**

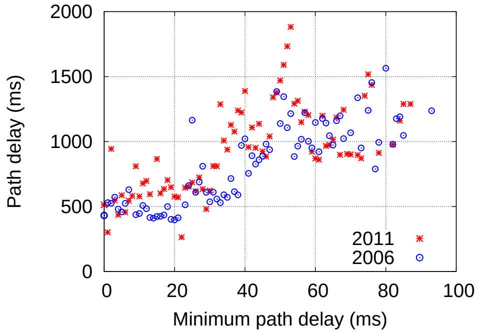


**Figure 11. Correlation between path delay and minimum path delay.**

### 3.2.5. Correlations between metrics

We also analyze the correlation between some metrics to further investigate the recent Internet evolution. In this analysis, we use the *Spearman* correlation coefficient, more indicated for non-linear relationships as it basically evaluates correlations between ranked values, thus also better dealing with outlier values.

Figure 9 shows the scatter graph correlating the path delay and its size (in number of hops). We clear note that path delay is highly correlated to path sizes lower than 20 for both years we analyze. We find a $0.8$ Spearman correlation coefficient for 2006 data and $0.88$ for 2011.

For path sizes larger than 20 hops, we observe that the latency remains roughly stable. This may occur because in large paths, larger than 20 hops, a small number of paths dominate the latency and thus, increasing the path with a couple of low latency links does not impact the overall latency. Moreover, this portion of the correlation (paths larger than 20 hops) represents a very small amount of paths. Remember that, according to Figure 2, the number of paths larger than 20 hops can be negligible.

We also correlated the geographic dispersion with the path delay dispersion. Here, we define the path delay dispersion as the ratio between the measured path delay and the minimum delay (i.e. only propagation delay) we would expect in this path using a fiber optic link. Following [Percacci and Vespignani 2003], digital information travels along fiber optic cables at almost exactly 2/3 the speed of light in a vacuum. So we can estimate the total delay using a direct fiber optic cable connecting two end-points. Figure 10 shows that path delay dispersion and geographic dispersion are highly correlated, especially to the major portion of paths (with low geographic dispersion values). The Spearman correlation coefficient in this case is $0.86$ for 2006 and is $0.88$ for 2011.

Finally, we analyze the correlation between the path delay and the minimum delay we expect in a fiber optic link. Figure 11 shows that, for both 2006 and 2011, the correlation seems to follow an exponential curve when path delay values are lower than $50ms$. Shorter paths may have their links in local fiber or cable links. In the opposite way, longer paths have a sort of links which may explain the bad correlation we found in longer delayed paths. We believe that these longer delayed paths are bounded by a small number of links with very large delays and, as consequence, the total path delay is not proportional to the minimum delay we would expect. We found better correlation for 2006, with Spearman coefficient about $0.78$ and a median correlation to 2011 paths, with Spearman coefficient about $0.67$.

## 4. Related Work

The characterization of Internet evolution and its behavior attracts research efforts for the last decade [Pastor-Satorras et al. 2001, Zhou 2006, Oliveira et al. 2007, Haddadi et al. 2008, Borgnat et al. 2009, Edwards et al. 2012]. Most of these works focus on how the Internet changes on the AS-level. Moreover, they mainly characterizes topological properties, lacking on end-to-end measures.

In particular, Siganos et al. [Siganos et al. 2002] showed in 2002 that the growth of the number of both nodes and edges at the AS-level has been exponential. This finding, with the observation that the Internet follows a "small world" structure, indicates a network densification. More recently, Dhamdhere and Dovrolis [Dhamdhere and Dovrolis 2011], almost a decade later after the work of Siganos et al., showed that the AS-level network growth is now linear. They also note that global economy and AS particular interests have strong influence in the topology dynamics. A key aspect, but rather unexplored in these works, is how those topological changes affect path diversity.

Recent works by Edwards et al. [Edwards et al. 2012] and Lee et al. [Lee et al. 2010] also present results that show apparent worsening of the network efficiency. The particular case of the increase in the perceived network delay, also observed in this paper, might be related to the *bufferbloat* phenomenon, recently investigated by Gettys et al. [Gettys and Nichols 2012]. Whiteaker et al. [Whiteaker et al. 2011] also show that the increasing delays in the network may be a consequence of the increasing number of virtualized systems. The heavy network usage from virtual machines can introduce delays as high as 100 ms.

## 5. Conclusions

In this work, we characterize and analyze the recent end-to-end connectivity evolution of the Internet. We compare key end-to-end performance metrics from two distinct periods separated by five years. We analyze data from 2006 and 2011, including the end-to-end path latencies, path number of hops, path geographic dispersion, and path diversity. Understanding this recent evolution in Internet connectivity, as well as its associated trends, is important in particular to inter-networking link planning and to better adjustment of routing protocols. Moreover, the findings we provide may be useful to support insights to foster the development of new protocols, especially routing protocols.

We found some counterintuitive results indicating that Internet performance is getting worse. For instance, the average path length distribution slightly changes from 2006 to 2011. During this period, however, we would expect networking investments (new equipments and links) to provide shorter paths currently.

Our analysis also shows that the end-to-end delay became clearly worse. Once again, new equipments and links should provide a better connectivity with faster links, resulting in an improved user experience. Nevertheless, in this case we just show the opposite expected result with delays experienced by user increasing. We conjecture that new Internet equipments are not properly configured to use active queue management. This leads to a recently identified phenomenon known as *bufferbloat* [Gettys and Nichols 2012], where the excess of packet buffering in the network causes high latency and jitter.

Furthermore, we show that path diversity decreased, and similarly, distinct paths became slightly more similar. This result has a direct impact on routing algorithms that try to explore the path diversity to be more fault-tolerant. This result indicates that Internet paths are growing around the minimum path, and as a consequence, the new links impact is marginal on the overall difference between two alternative end-to-end paths.

Finally, we found that the delay is highly correlated to the path number of hops in the larger portion of end-to-end paths. The geographic dispersion has a high correlation with the path delay, especially to low values in geographic dispersion (which we found in almost all paths). We also note that path delay has an exponential correlation pattern with the minimum path delay expected (for delays bellow 50 ms). These correlation patterns for the large portion of paths indicate that we can infer metric values using a quite simple metric, such as the round trip time.

As future work, we intend to investigate deeper the *bufferbloat* phenomenon and its relation to the path delay we found in the network traces we analyzed.

## References

Borgnat, P., Dewaele, G., Fukuda, K., Abry, P., and Cho, K. (2009). Seven Years and One Day: Sketching the Evolution of Internet Traffic. In *Proc. of the IEEE INFOCOM*.

Dhamdhere, A. and Dovrolis, C. (2008). Ten Years in the Evolution of the Internet Ecosystem. In *Proc. of the Internet Measurement Conference (IMC)*.

Dhamdhere, A. and Dovrolis, C. (2011). Twelve years in the evolution of the internet ecosystem. *IEEE/ACM Transactions on Networking (TON)*, 19(5).

Edwards, B., Hofmeyr, S., and Stelle, G. (2012). Internet Topology over Time. *Arxiv preprint arXiv:*.

Gettys, J. and Nichols, K. (2012). Bufferbloat: dark buffers in the Internet. *Comm. of the ACM*, 55(1).

Haddadi, H., Uhlig, S., Andrew Moore, A., Mortier, R., and Rio, M. (2008). Modeling Internet Topology Dynamics. *ACM SIGCOMM Computer Communication Review*, 38(2).

Lee, D., Cho, K., Iannaccone, G., and Moon, S. (2010). Has Internet Delay gotten Better or Worse? In *Proc. of the 5th International Conference on Future Internet Technologies*.

Oliveira, R., Zhang, B., and Zhang, L. (2007). Observing the evolution of internet as topology. *ACM SIGCOMM Computer Communication Review*, 37(4).

Pastor-Satorras, R., Vázquez, A., and Vespignani, A. (2001). Dynamical and Correlation Properties of the Internet. *Physical Review Letters*, 87(25).

Percacci, R. and Vespignani, A. (2003). Scale-free Behavior of the Internet Global Performance. *Eur. Phys. J. B*, 32:411–414.

Siganos, G., Faloutsos, M., and Faloutsos, C. (2002). The Evolution of the Internet: Topology and Routing. Technical report, Univ. of California, Riverside.

Whiteaker, J., Schneider, F., and Teixeira, R. (2011). Explaining Packet Delays Under Virtualization. *ACM SIGCOMM Computer Communication Review*, 41(1):38–44.

Zhou, S. (2006). Understanding the Evolution Dynamics of Internet Topology. *Phys. Review E*, 74(1).