Five Years at the Edge: Watching Internet from the ISP Network

Martino Trevisan, Danilo Giordano, Idilio Drago, Maurizio M. Munafò, Marco Mellia
Politecnico di Torino
first.last@polito.it

***Abstract*—The Internet and the way people use it are constantly changing. Knowing traffic is crucial for operating the network, understanding users' needs, and ultimately improving applications. Here, we provide an in-depth longitudinal view of Internet traffic during 5 years (from 2013 to 2017). We take the point of the view of a national-wide ISP and analyze rich flow-level measurements to pinpoint and quantify changes. We observe the traffic, both from a point of view of users and services. We show that an ordinary broadband subscriber downloaded in 2017 more than twice as much as they used to do 5 years before. Bandwidth hungry video services drove this change at the beginning, while recently social messaging applications contribute to increase of data consumption. We study how protocols and service infrastructures evolve over time, highlighting events that may challenge traffic management policies. In the rush to bring servers closer and closer to users, we witness the birth of the sub-millisecond Internet, with caches located directly at ISP edges. The picture we take shows a lively Internet that always evolves and suddenly changes. To support new analyses, we make anonymized data available at https://smartdata.polito.it/five-years-at-the-edge/.**

***Index Terms*—Passive Measurements; Broadband Characterization, Longitudinal Traffic Analysis, Service Usage.**

I. Introduction

Network measurements are the key means to gather information about the overall status of the Internet, identify eventual issues, and ultimately understand how the network is evolving [34], [47], [41]. However, having a long-term picture on the Internet evolution is a rather challenging task. Continuous efforts on monitoring such trends from measurements are rare [7], and understandably limited in coverage, duration and location (e.g., see [6], [15], [18], [19], [31]). In addition we need different prespectives to maintain and complement our understanding of the Internet ecosystem.

In this paper,[1] we offer a longitudinal view on the Internet usage and evolution, covering a 5-year long interval (2013–2017). We rely on a humongous amount of data collected from a nation-wide Internet Service Provider (ISP) infrastructure. We instrument some of the ISP aggregation links with passive monitoring probes. By observing packets flowing on links, our probes extract detailed per flow information, that we collect and store on a centralized data lake. Keeping the pace with Internet evolution during 5 years is *per se* a challenging task. We rely on custom software (Tstat [44]) that have been constantly updated during the monitoring period to account for and report information about new protocols and services.

[1] An early version of this work has appeared in [45].

Technically, we follow a well-established approach (Section II). Passive measurements are popular among researchers since early 2000 [3], [13], with current tools able to process several tens of Gb/s on commodity hardware [35]. Extracting information from packets is possible thanks to Deep Packet Inspection (DPI) techniques [1], while the availability of big data solutions [14], [48] makes it possible to store and process large volumes of traffic with unprecedented parallelism. Here, we dive into this data, depicting trends, highlighting changes and observing sudden infrastructure upgrades.

First we offer an overview of users' habits over 5 years, describing the traffic consumption posed by broadband customers to the ISP (Section III). Next, we turn our attention to traffic generated by individual services. We quantify the rise (and death) of services considering traffic volume as well as popularity among customers (Section IV). We analyze protocol usage and episodes of changes in services that result on unpredictable traffic (Section V). Finally we study how changes in the infrastructure impact the ISP network (Section VI).

Our analysis leads to new insights, confirms well known trends and quantifies interesting aspects about Internet traffic in a broadband edge network, summarized in Section VII. Some highlights are as follows:

- The traffic per broadband customer has increased at a constant rate over the years (2013–2017), with a growth of heavy users, who exchange tens of GB per day;
- Compared to ADSL customers, the larger capacity offered to FTTH customers has a moderate impact on data consumption;
- We witness the (slow) migration of services to new protocols (e.g., HTTPS) and several sudden changes performed by over-the-top Internet companies. We report some cases that required the ISP to act promptly with traffic engineering and troubleshooting;
- We quantify well known Internet trends: video content still drives bandwidth demands; peer-to-peer traffic, while heading to insignificance, persists with few loyal users; traffic from mobile devices connected via WiFi is prominent at home networks;
- We observe the rise of new "elephant" services, in particular social networks such as Instagram, accessed from mobile phones at home. We find that Instagram traffic is comparable to video-on-demand services, such as Netflix or YouTube.
- We testify the infrastructure growth of popular services, and show how their servers are getting closer to users.

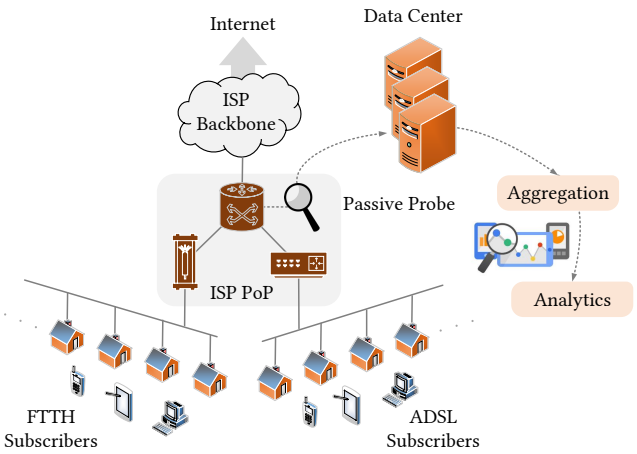


Fig. 1: Measurement infrastructure and processing steps.

In the most extreme case, we identify the deployment of caches at the first aggregation point at the ISP to reduce the latency to reach contents.

To foster new analyses beyond what is presented here, we make anonymized data available to the community. Monthly snapshots are available online. Other snapshots can be accessed following instructions found at the our website.[2]

Despite our dataset being limited to one country and focused on broadband Internet connections (thus missing mobile access networks), we believe the information we offer is key to understand trends and inform researchers and practitioners about recent changes on Internet infrastructure and users' behavior.

II. Measurement Methodology

We next describe the monitoring architecture, measurement methodology and tools we use to collect the data.

A. Measurement architecture

We build on data collected by the passive monitoring infrastructure of a nation-wide ISP in Italy. It operates a backbone tier-2 network, connected to hundreds of customer and peering ASes and a handful of provider ASes.

The passive monitoring infrastructure captures and analyses in real-time traffic from vantage points located at the edge of the ISP network. A schematic view of the infrastructure is depicted in Figure 1. We process traffic directly in the ISP Points-of-Presence (PoPs). Exploiting router span ports or optical splitters (depending on the link rates), the traffic is mirrored to the monitoring probes. Both uplink and downlink streams are processed by the probes. Since probes are deployed in the first level of aggregation of the ISP, no traffic sampling is performed. Customers are assigned fixed IP addresses, that the probes immediately anonymize in a consistent way.

Each probe is equipped with multiple high-end network interfaces. Packets are captured using the Intel Data Plane Development Kit (DPDK) [25] that guarantees line-rate capture even for multiple 10 Gbit/s links. Traffic is then processed by our custom-made passive traffic analyzer, called Tstat [44].

Each probe exports only flow records, i.e., a single entry for each TCP/UDP stream with per-flow statistics. Streams are expired either by the observation of particular packets (e.g., TCP packets with RST flag set) or by the default Tstat timeouts [44]. Each record contains classical fields on flow monitoring [23], such as IP addresses, port numbers, packet- and byte-wise counters. Advanced analyzers extract some few fields from packet payloads, such as information seen in the Application-Layer Protocol Negotiation (ALPN) fields of TLS handshakes, which allows us to identify HTTP/2 and SPDY flows, and fields from QUIC public headers.

Tstat exports the *domain name* of contacted servers, which is essential for service identification [43]. Tstat first searches the information in HTTP `Host:` headers and in the TLS Server Name Indication (SNI) within TLS Client Hello messages. For HTTP/TLS flows missing such fields and for other protocols, Tstat exports the hostname the client resolved via DNS queries prior to open the flow. This is achieved by caching DNS traffic directed to any DNS resolver. Once a flow without SNI or HTTP `Host:` is observed, tstat searches for the last query performed by the same client that resulted in the contacted server IP address. This mechanism, called DN-Hunter, is explained in details in [5], where it is shown that the association is correct for more than 90% of the flows.

B. Characteristics of the dataset

Among the vantage points, here we consider the traffic of two PoPs, covering more than 10 000 ADSL and 5 000 Fiber-To-The-Home (FTTH) subscribers, all located in the same city in Italy. ADSL downlink capacity varies from 4 Mbit/s up to 20 Mbit/s, with uplink limited to 1 Mb/s. FTTH subscribers enjoy 100 Mb/s downlink, and 10 Mbit/s uplink. Each subscription refers to an ADSL or FTTH installation, where users' devices (PCs, smartphones, tablets, smart TVs, etc.) reach the Internet via WiFi and Ethernet through a home gateway. ADSL customers are mainly residential customers (i.e., households), whereas a small but significant number of business customers exist among the FTTH customers.

During the 5 years of measurements we observed a steady reduction on the number of active ADSL subscribers and an increase in FTTH installations. The ISP has confirmed these trends are due to churning and technology upgrades. To compensate for such changes, we will report statistics aggregating measurements and normalizing numbers according to the number of active subscribers per day.

C. Data storage and processing

Flow records are created, anonymized and stored on the probe local disks. Daily, logs are copied into a long-term storage in a centralized data center and discarded from the probes.

The considered dataset covers 5 years of measurements, totaling 31.9 TB of compressed and anonymized flow logs (around 247 billion flow records). To process this data, we use a Hadoop-based cluster running Apache Spark. This structure

[2]https://smartdata.polito.it/five-years-at-the-edge/

TABLE I: Examples of domain-to-service associations.

Domain	Service
facebook.com	Facebook
fbcdn.com	Facebook
^fbstatic-[a-z].akamaihd.net$ (RegExp)	Facebook
netflix.com	Netflix
nflxvideo.net	Netflix

allows us both to continuously run predefined analytics, as well as to run specific queries on historical collections.

Our methodology follows a two-stage approach: First, data is aggregated on a per day basis; Second, advanced analytics and visualizations are computed. In the aggregation stage, queries compute per-day and per-subscription aggregates about traffic consumption, protocol usage, and contacted services.

Special attention is needed to identify the services used by subscribers. Content providers are known to rely on large infrastructure and/or Content Delivery Networks (CDNs), which make the association between flow records and services tricky. We rely mostly on server hostname for this step. Examples of the association domain-service are provided in Table I. Flexible matching based on regular expressions are used.[3] Along the years, our team has *continuously* monitored the most common server hostname seen in the network, maintaining the list of domains associated with the services of interest. For ambiguous cases [43], e.g., domains used by multiple services, we rely on heuristics, mostly based on traffic volumes, to decide whether a subscriber actually contacted a particular service (see Section IV-A). This methodology allows on-the-fly and historical classification of services.

D. Measuring the distance to servers

To analyze content placement and providers' infrastructure (Section VI), we rely on two sources.

First, we rely on the estimation of RTT provided by Tstat for TCP flows [33]. It matches acknowledgements with TCP data segments, registering the time from the observation of the TCP segment and its acknowledgment. For each flow, Tstat exports the minimum, average and maximum RTT estimation, as well as the number of RTT samples. Notice that this metric represents the RTT from the probe to servers, missing the delay from clients to the probes. Thus, in our deployment we ignore the access delay, since probes are deployed at the ISP's PoPs.

Second, we rely on BGP measurements for (i) analyzing the path taken by packets to reach the ISP from external sources, and (ii) evaluating the traffic entering the ISP network from *customer*, *peer* and *provider* ASes. We focus on incoming traffic as it represents the majority of the volume for the broadband customers.

We map external IP addresses to the corresponding ASes using RouteViews [10]. Then, we determine the paths taken by packets to reach the ISP.

We first rebuild the AS topology using CAIDA's Relationship dataset [9]. This dataset determines the relationships between ASes classified as *customer-provider* or *peering*. We then apply the *valley free* rules [30] to obtain a list of plausible paths reaching the ISP from the ASes hosting the external IP addresses seen in our traces. This methodology is always applicable, but since it is based on heuristics it may report nonexistent paths. However, it follows standard BGP practices of route selection.

To refine the paths we rebuild the AS topology using BGPStream [37]. BGPStream provides fine-grained routing information, but only for a subset of sources. We use BGPStream to extract all Internet paths seen on the 15th day of each month, using Routing Information Bases (RIBs) observed from 380 vantage points from the Route Views and other 600 sources from RIPE's Routing Information Service (RIS). Then, we keep only paths having as destination the ISP AS.

Finally, we merge the two obtained topologies. We reexamine all paths discovered through CAIDA's Relationship dataset. In case BGPStream data is available for an AS path (or part of it), we discard CAIDA's Relationship (sub-)path and use BGPStream's one. Again, we enforce the *valley free* rules [30].

For each external IP address, we calculate the shortest path to the studied ISP. If multiple shortest paths are available, we give priority to paths starting with a *customer* or *peer* AS, rather than a *provider* AS. The rationale is that the AS would not pay for such traffic. We save also the *ingress* AS, i.e., the last hop before the studied ISP, to determine whether the traffic reaches the ISP through *customer*, *peer* or *provider* ASes. Again, if more than one shortest path is available, multiple ingress points/relations might exist. Such ambiguous relations are marked as *Not Available* (NA).

This methodology determines the shortest path and the AS relation of the ingress point for each flow. Based on these metrics, we study how far (in terms of AS path length) the content is deployed from users, and whether this distance has varied throughout years. Moreover it sheds some light on the relationship between the ISP and content sources.

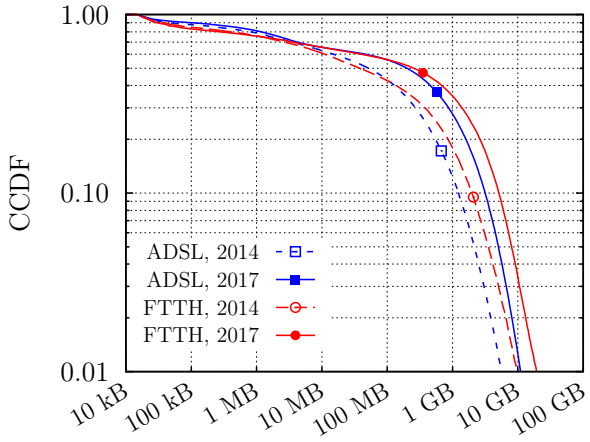
E. Challenges in long-term measurements

Several challenges arise when handling a large-scale measurement infrastructure. Network probes are the most likely point of failure, as they are subject to a continuous and high workload. During the period considered in the paper, probes suffered some outages, lasting from few hours up to some months (when severe hardware issues arose). As such, the results we present have missing data for those periods. The FTTH probe in particular has suffered outages in 2013. For this reason, when comparing FTTH and ADSL customers we will focus on the period from 2014 to 2017.

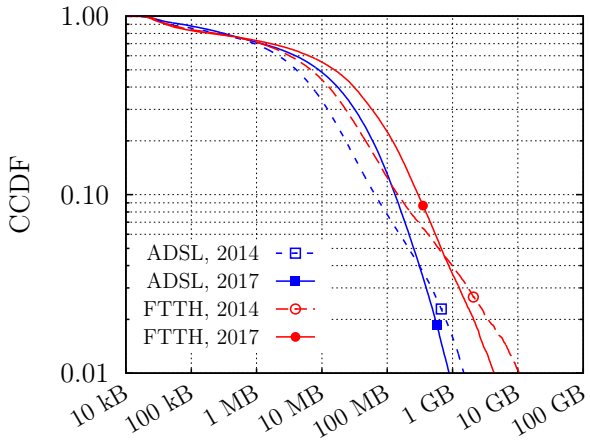
A second issue arises from the evolution of network protocols and service infrastructure. Large content providers have the power of suddenly deploying new protocols leaving passive monitors and ISPs with few or no documentation to handle them. We incurred several cases, and report our experience in addressing them.

Third, the domain-to-service associations need to be continuously updated. Also in this case, there is no public information to support this operation, so our team has to manually

[3]The full list of regular expressions used to classify services are found at https://smartdata.polito.it/five-years-at-the-edge/.



(a) Download volumes



(b) Upload volumes

Fig. 2: Daily traffic per active customer for Apr 2014 and 2017.

define and update rules, often running active experiments to observe new patterns.

At last, users' privacy must be preserved. For this, we carefully limit the collected information and always consider only aggregated statistics. Customers' IP addresses and server names are the most privacy-sensitive information being collected. The former gets immediately anonymized by probes, while the latter is used to derive aggregate statistics on per-service basis. Importantly, all data collection is approved and supervised by the responsible teams in the ISP.

III. Subscriber's traffic consumption

We first characterize the amount of traffic consumed by subscribers. For the results that follow, we consider only *active subscribers*. Subscribers are considered *active* in a given day if they have generated at least 10 flows, downloaded more than 15 kB and uploaded more than 5 kB. These thresholds have been determined by inspecting the distributions of daily traffic per subscriber, visually searching for knee points (see Appendix A). This criterion lets us filter those cases where only background traffic is present, e.g., generated by the access gateway, or by incoming traffic (due to, e.g., port scans). On average we observe about 80% subscribers are active in a day with respect to the total number of subscribers observed in the whole trace.

Notice that these percentages are actually a lower-bound given churning (see Section II-A).

A. How much you eat: Consumption per day

Figure 2 depicts the empirical Complementary Cumulative Distribution Function (CCDF) of daily traffic consumption of active subscribers. In other words, for each day, we compute the overall traffic each active subscriber exchanges, and report the CCDF of all measurements. We focus on April 2014 and April 2017. Figure 2 depicts CCDFs separately per access-link technology and down/up links. Notice the log scales.

Observe the bimodal shape of the distribution. In about 50% of days, subscribers download (upload) less than 100 MB (10 MB) – i.e., days of light usage. However, a heavy tail is present. For more than 10% of the days, subscribers download (upload) more than 1 GB (100 MB) – i.e., days of heavy usage. Manual inspection shows that many different subscribers present days of heavy usage, often alternating between days of light and heavy usage.

Comparing 2014 (dashed lines) with 2017 (solid lines), we notice a general increase in daily traffic consumption. The median values have increased by a factor 2 for both ADSL and FTTH installations, and for both upload and download. This behavior highlights an increasing trend in average per subscriber traffic volume, that we examine more in depth later in this section.

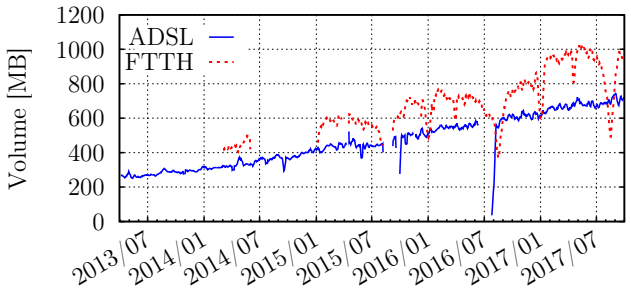
We observe no difference for the days of light usage when contrasting ADSL (blue curves) and FTTH installations (red curves). Instead, during heavy usage days, FTTH subscribers download about 25% more data than ADSL subscribers – a moderate increase given they enjoy 5-20 times higher capacity. The differences are higher considering upload traffic: ADSL subscribers are indeed bottlenecked by the 1 Mb/s uplink, thus FTTH subscribers upload twice as much per day.

At last, we witness an interesting effect in upload traffic: Even if traffic volume increases in median between 2014 and 2017, the tail of the distributions in Figure 2b decreases. Notice the clearly visible bump in the tails present in 2014, which disappeared in 2017. A deeper analysis on per-subscriber traffic distribution is provided in Appendix B. This trend is rooted in the decline of Peer-To-Peer (P2P) traffic, both in volume and popularity, as we will show in Section IV.

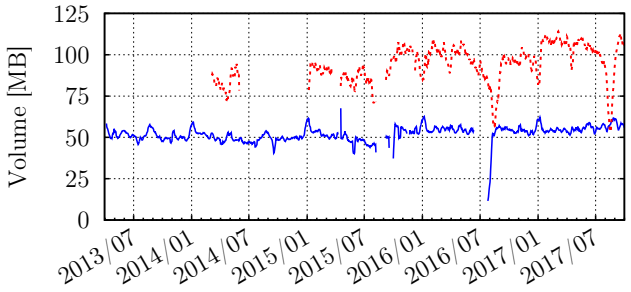
B. Eager and Eager: Trends on traffic consumption

Figure 3 illustrates the per-subscriber traffic consumption over time. The x-axis spans over the 54 months of the dataset, while y-axis shows the average byte consumption over monitored subscriptions, separately per access technology and down/up link. We here show only average trends, with more information (e.g., percentiles) provided in Appendix B.

Curves in the figure contain interruptions caused by outages in the probes. Note in particular the outages in FTTH probe in 2013. FTTH figures are noisier than the ADSL ones because of the smaller numbers of FTTH customers. Some drops in

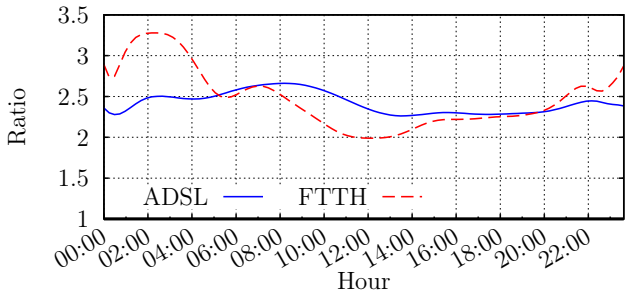


(a) Download volume

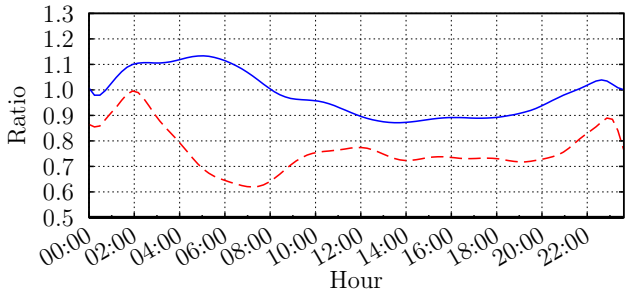


(b) Upload volume

Fig. 3: Average per-subscription daily traffic.



(a) Download



(b) Upload

Fig. 4: Ratio of download and upload traffic per hour of the day, considering April 2017 and April 2014.

FTTH curves are visible during summer and holiday breaks, due to the small number of customers and their usage profiles, e.g., the presence of business customers who stop using the network on holidays.

Considering the average amount of data daily downloaded, illustrated by Figure 3a, a clear increasing trend emerges. For ADSL subscribers, average daily traffic increased at a constant rate – from 300 MB per day in 2013 up to 700 MB in late 2017. FTTH subscribers consume on average 25% more traffic, topping to 1 GB per day on average in 2017. Interesting, very similar slow increasing trends have been reported 10 years ago [12].

When considering uploads (Figure 3b), we confirm that the higher uplink capacity lets FTTH users upload more with respect to ADSL. The latter are bottlenecked by the limited upload capacity and thus the average amount of data remains constant. FTTH subscribers show instead a modest increase in average uploaded traffic over time. This increase is due to two factors. At the one hand, P2P uploads have decreased significantly in recent years. On the other hand, this decrease has been compensated by a significant increase in the upload of user-generated content to the cloud, including backups to cloud storage services (e.g., iCloud or Dropbox) as well as to social networks and video providers (e.g., YouTube and Instagram).

To check whether the increase observed in Figure 3 is homogeneous during the hours of the day, we consider the download and upload volumes in each 10 minute-long time interval. We then average all values seen for the same time bin in all days of a month. At last we compute the ratio between days of April 2017 and April 2014.

Figure 4 shows results (curves are smoothed using a Bezier interpolation). Observing first the downloads (Figure 4a), we confirm that the average amount of traffic consumed in 2017 is more than 2 times larger than 2014. More interesting, the increase is higher during late night hours. Manual inspection suggests that the increase in late night traffic is a consequence of diverse factors, such as the automatic download of updates of smartphone apps and other machine-generated traffic, such as from home IoT devices. FTTH subscribers exhibit a higher increase also during prime time, which we confirm to be associated to the consumption of video streaming content.

When considering uploads (Figure 4b), we again see significant differences throughout the day, even if the overall figure has remained almost constant. Observe how the upload volume has actually decreased during daytime for both FTTH and ADSL subscribers, whereas night traffic has increased. The former can be associated to the decline of P2P, whereas the latter is partly driven by automatic backups performed by some applications – e.g., WhatsApp, performing backups at 2AM.

IV. Evolution on the Usage of Services

A. *Give me that: Service popularity*

The changes in the per-subscriber traffic volume can be due to changes in the users' habits (e.g., people using different services), or changes in the services (e.g., high definition videos being automatically served). In this section, we analyze in details how popular and bandwidth demanding services evolved throughout years. We again focus on *active subscribers*, observing the fraction of them that accessed a given *service* on a daily basis. In the analysis that follows, we focus on downloaded volume only.

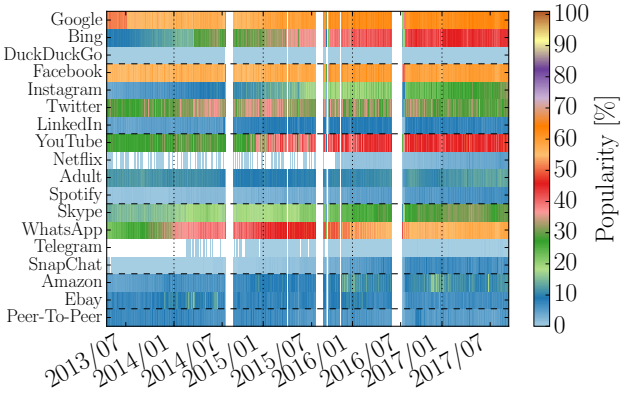


Fig. 5: Popularity for selected services over time.

Notice that selecting those subscribers that contacted a service is not trivial. Indeed popular services may be *unintentionally* contacted by users. Consider for example Facebook. Its social buttons are embedded in websites and generate traffic to the same Facebook domains as an intentional access to `facebook.com` services. To coarsely distinguish these cases, we have inspected the distribution of daily traffic per subscriber for each considered service, setting per-service thresholds to separate (i) subscribers with at least one intentional visit to the target service (moderate to large traffic volumes), and (ii) subscribers which unintentional contacted domains due third party objects (negligible volumes).

We first provide a coarse picture about service popularity over time. Figure 5 shows per-day percentage of active subscribers that access popular services. We show the ADSL data only, since FTTH results in similar figures. Similar figure for downloaded bytes is reported in Appendix A. The multi-color palette highlights changes in the popularity of services, which are coarsely sorted by type. For instance, Google Search is accessed regularly by about 60% of active subscribers, and this pattern is rather constant over time.[4] On the contrary, Bing shows a constant growth, moving from less than 15% to about 45% of active subscribers that contacted it at least one time per day in 2017. This pattern is likely a consequence of Windows telemetry which uses `bing.com` domains. DuckDuckGo, a privacy respecting search engine, is used only by few tens of subscribers (less than 0.3% of population).

Overall, we observe a continuously changing picture, with services showing increase in popularity, some of which with remarkable growth, while others that struggle to gain grounds. Next, we dive into some interesting cases.

B. The downfall of Peer-To-Peer – finally

It is no news that P2P is no longer among the preferred means to download content. Here we quantify this phenomenon showing the popularity of P2P applications over the years. Figure 6a details the percentage of active subscribers using a P2P service (Bittorrent, eMule and variants) (top plot) and the average P2P traffic volume per subscriber (bottom plot). We still observe a hardcore group of subscribers that exchange about 400 MB of P2P data daily. At end of 2016 the traffic volume they generate starts to decrease. Interestingly, FTTH subscribers start abandoning P2P applications earlier in terms of volume. Based on findings of previous studies [21], [31], a conjecture to explain this decline is that the availability of cheap, easy and legal platforms to access content is finally contributing to the downfall of P2P. In the following we explore this conjecture.

[4]Some fluctuations are due to changes in Google domains that took time to be identified and updated in the probes.

C. The usual suspects: YouTube and Netflix

We consider popular video streaming services. Figure 6b shows the percentage of active subscribers accessing Netflix (top) and the average per-subscriber daily traffic (bottom). Netflix has gained momentum since the day it started operating in Italy. FTTH subscribers have been eager to adopt it, with about 10% of the ISP customers using it on a daily basis at the end of 2017. Considering weekly statistics (not shown in the figure), more than 18% (12%) of FTTH (ADSL) subscribers access Netflix at least once in 2017.

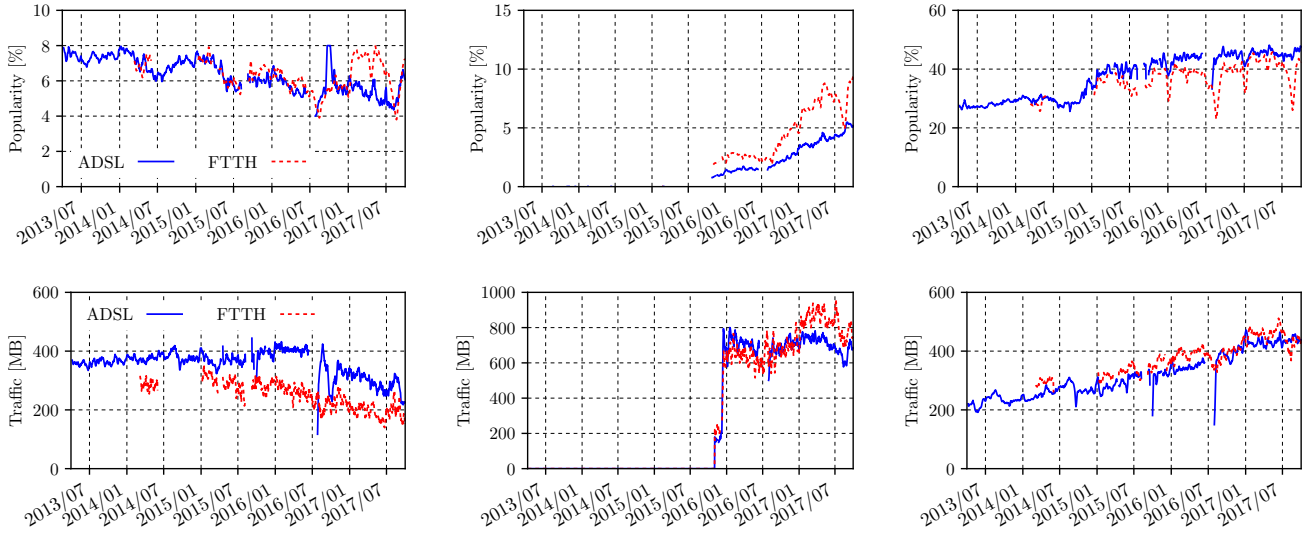
Considering the amount of traffic they consume (bottom plot), we see no major differences between ADSL and FTTH subscribers up to end of 2016. Since October 2016, Netflix started offering Ultra HD content. This is reflected into each active FTTH subscriber downloading close to 1 GB of content on average per day. Such a high traffic volume well justifies the large scale content delivery infrastructure of Netflix [8]. ADSL subscribers instead cannot enjoy Ultra HD content, or are not willing to pay the extra fee.

Next, we focus on YouTube (Figure 6c). The figure shows a consolidated service, accessed regularly by users, who are consuming more and more content: more than 40% of active subscribers access it daily, and download more than 400 MB (about half of Netflix volume per subscriber). Interestingly, no differences are observed between ADSL and FTTH subscribers – hinting that YouTube video works similarly on FTTH and ADSL.

D. New elephants in the room: Social messaging applications

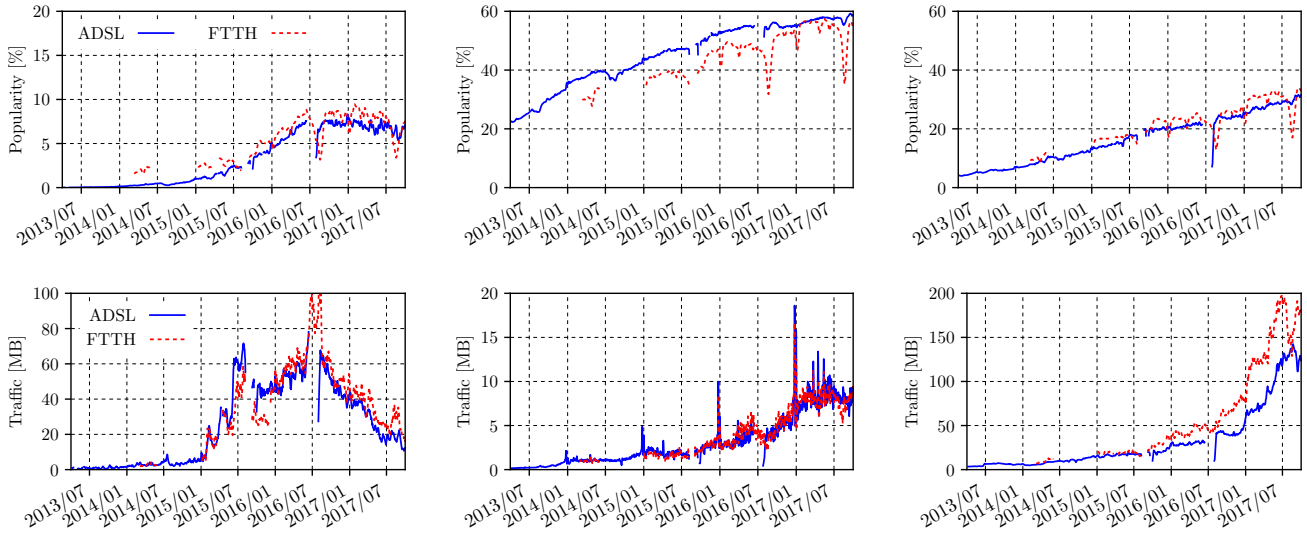
We now study usage patterns for social messaging applications, namely SnapChat, WhatsApp and Instagram. All are popular applications accessed mostly from smartphones, whose traffic we observe once connected via WiFi from home. As before, we consider popularity and daily download traffic consumption per active subscriber (recall Section IV-A). Results are depicted in top and bottom plots in Figure 7.

Interesting trends emerge in the rise and fall of social networking apps. Observe first Snapchat (Figure 7a). It enjoyed a period of notoriety starting from 2015, topping in 2016 when it was adopted by around 8% of subscribers. Each active subscriber used to download up to 100 MB of data daily! Starting from 2017, the volume of data starts to decrease, with active subscribers that nowadays download less than 20 MB per day. Popularity is mostly unaffected, suggesting that people keep having the Snapchat app, but seldom use it.



(a) Peer-To-Peer (b) Netflix (c) YouTube

Fig. 6: Popularity (top) and volumes (bottom) for P2P and 2 popular video streaming services.



(a) SnapChat (b) WhatsApp (c) Instagram

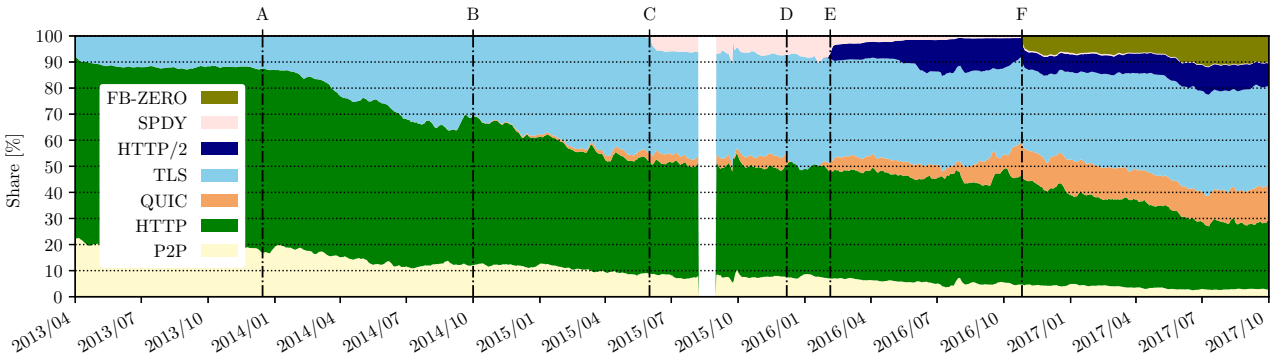
Fig. 7: Popularity (top) and volumes (bottom) for 3 popular social messaging services.

The decline of SnapChat coincides with the growth of other social apps. See WhatsApp in Figure 7b: Its popularity is indisputable, with a steady growth in adopters that has almost reached saturation. Observe instead the growth in daily volume per active subscriber. Each subscriber downloads around 10 MB daily, pointing to the intensive use of the app for sharing multimedia content. Note also the large peaks in the figure, corresponding to Christmas and New Year's Eve, when people exchange wishes using WhatsApp.

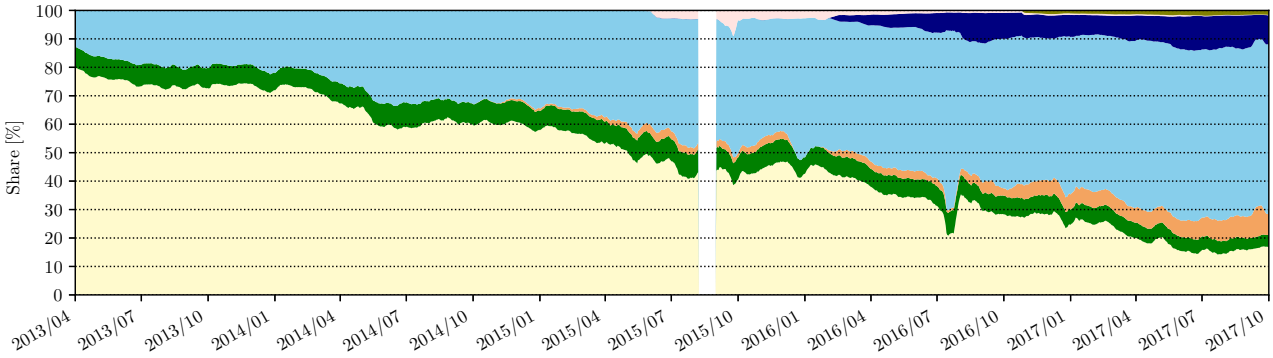
Finally, consider Instagram (Figure 7c). We see a constant growth in popularity and, more impressive, a massive growth in download traffic volumes. Each active subscriber downloads on average 200 MB and 120 MB per day, for FTTH and ADSL respectively. This is almost a quarter of the traffic of the active customers contacting Netflix. Recalling that Instagram, Snapchat and WhatsApp are predominantly used from mobile terminals, these figures point to a shift on traffic of broadband users, with mobile terminals taking a predominant role even when people are at home. This is confirmed in [46].

V. Web trends, and surprises

We next study how protocol usage varied across 5 years. We show events associated with the slow migration of services towards standard protocols, and sudden relevant changes on



(a) Download



(b) Upload

Fig. 8: Protocol breakdown over 5 years. Sudden changes and custom protocols deployment in the wild are highlighted.

the traffic caused by experiments of big players with custom protocols.

In its early life, the Internet was predominantly plain Hyper Text Transfer Protocol (HTTP) and P2P traffic. It is by now known that most of the Internet traffic is running on encrypted web protocols [36], first with the deployment of HTTPS, followed by the push towards HTTP/2 [4] (which is always carried over TLS) and more recently QUIC [28]. We here document to what extent these protocols have been adopted.

Figure 8 answers this question by summarizing the ADSL data, with the omitted FTTH figures showing similar trends. Figure 8a shows the download traffic share of protocols over time. In 2013, only the three "classic" protocols were observed: The majority of the traffic was served by plain-text HTTP, around 10% of the traffic was due to TLS/HTTPS and the remaining 20% of the traffic was due to P2P applications. In subsequent months, while P2P was showing its steady decline, several other notable changes happened, which are marked with letters in the figure:[5]

A) January 2014: YouTube starts serving video streams over HTTPS. The migration has taken Google several months during 2014, in which we can see a steady change in the mix of HTTP and HTTPS traffic. HTTPS share tops to around 40% at the end of 2014, and it is mainly driven by YouTube traffic.

B) October 2014: After announcing it in 2013, Google starts testing QUIC in the wild by deploying it on Chrome Browser. Web traffic carried by QUIC (over UDP) starts growing steadily.

C) June 2015: We update our probes to explicitly report SPDY protocol (previously generically labeled as HTTPS). We discover 10% of the traffic carried by an experimental protocol.

D) December 2015: Google disables QUIC for security issues [28]. Suddenly 8% of the traffic falls back to TCP and HTTPS/SPDY. Around a month after, the bug is fixed and QUIC is suddenly back.

E) February 2016: Google migrates traffic from SPDY to HTTP/2, slowly followed by other players.

F) November 2016: Facebook suddenly deploys "FB-Zero", a protocol with a custom 0-RTT modification of TLS, used by the Facebook mobile app only. Zero protocol would be announced only in Jan 2017.[6] Suddenly, 8% of the traffic moves to the new protocol. More than a half of Facebook traffic is carried by Zero, showing that its mobile app traffic surpassed website access, even for fixed ADSL installations.

At the end of 2017, HTTP traffic is down to around 25%, and HTTP/2 is slowly gaining momentum. QUIC and Zero together carry 20–25% of the traffic. Both were yet to be

[5]These events have been confirmed manually throughout the years while upgrading the software of our probes to keep-up with protocols evolution.

[6]https://goo.gl/vuQ1Jy

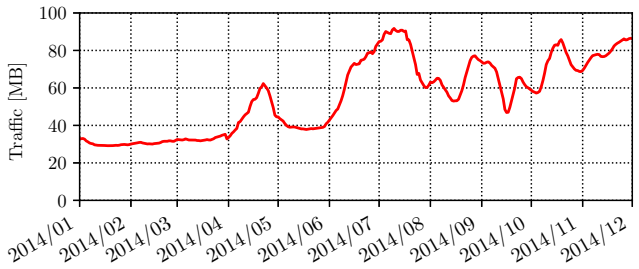


Fig. 9: Facebook average daily per-user traffic before and after automatic video play.

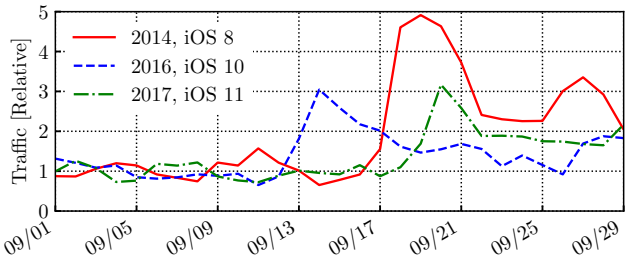


Fig. 10: Daily relative traffic to Apple's servers. The average traffic in the 1st week of each month is taken as reference.

standardized protocols, showing how giants like Google and Facebook are free to deploy experiments on the Internet, since they own both server and client applications. Such experiments may create issues for ISP network administrators, e.g., making network proxies and firewalls suddenly inefficient, or creating issues with home gateways.

Figure 8b complements the analysis showing upload traffic. Recall that the overall upload volume has remained mostly constant in this network throughout the years. Here we confirm and quantify the impact of the P2P decline in upload traffic. Observe how the traffic share of P2P has been replaced by HTTPS, and later on by HTTP/2 too. These two protocols are used to carry user-generated content in e.g., social networks, backups in cloud storage, etc.

A. Notable episodes

Here we report episodes that demonstrate how the traffic due to specific services can dramatically change in short time challenging traffic management at the ISP. We show (i) a sudden change in service setup with huge impact on bandwidth consumption; (ii) programmed software updates; (iii) services driving growth on late night traffic.

1) Facebook video autoplay: We illustrate an episode of sudden traffic changes. Around March/April 2014, Facebook started enabling video auto-play for its applications. The immediate effect on ISP traffic is striking. Figure 9 illustrates the daily average traffic per subscriber towards Facebook. Starting in March 2014 traffic has grown from around 35 MB to around 70 MB in a month. After an apparent stop in the deployment of the feature during May, Facebook enabled video auto-play again. In July, the daily traffic per subscriber was around 90 MB on average, 2.5 times higher than the rate observed in March 2014. This figure illustrates once more how the big players controlling client and server software can freely deploy changes in the Internet, complicating the planning and management of networks.

2) Scheduled elephants – iOS updates: Modern operating systems and mobile apps perform automatic updates to keep users' device safe. When updates are rolled out, a potentially large number of devices would trigger the download of new software. Here we report an example of the impact of such events on ISP networks: The release of Apple iOS updates.

Figure 10 reports the *relative variation* of the Apple traffic during September 2014, 2016 and 2017, when iOS updates have been released. The figure is built taking as reference the average traffic to Apple's servers in the first week of the respective months. It shows the relative variation of the daily traffic volume when compared to the reference week.

Notice how the traffic to Apple's servers increases during the update release periods. During the iOS 8 release in 2014, for example, we observe a five-fold increase in traffic, which lasted for a couple of days. Considering that the amount of traffic to Apple is already high (hundreds of GBs per day, e.g., due to iTunes and iCloud), the potential impact of such events in ISP networks is remarkable.

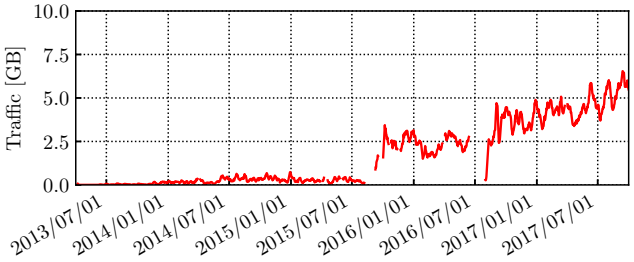


Fig. 11: Nightly Google Drive upload traffic.

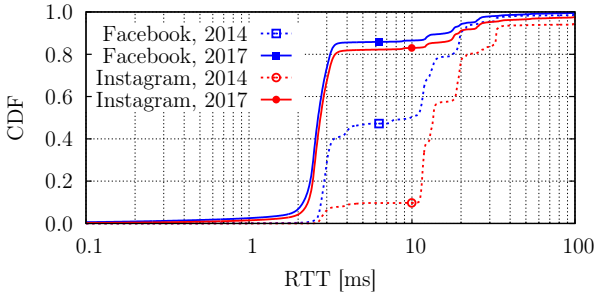
3) Night elephants – WhatsApp backups on Google Drive: The WhatsApp app offers users to perform automatic backups. It uses Google Drive for storing data, scheduling backups at 2 AM by default, and only performs the backup when the device is connected to WiFi.

Figure 11 shows the overall upload traffic to Google Drive during 1-hour intervals starting at 2 AM for the five years of capture. Notice how Google Drive's upload traffic used to be irrelevant before August 2015. Since then, it starts to grow steadily. By the end of 2017, Google Drive upload traffic at 2 AM was already over 5 GB.

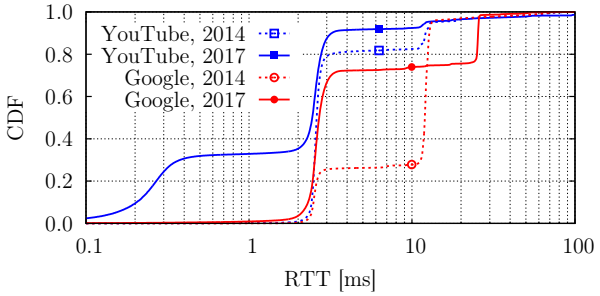
These numbers point to a change on night Internet traffic. While such traffic used to be dominated by P2P, night elephant flows are now observed towards cloud infrastructure.

VI. Where are my servers?

In the previous section we have reported both slow and sudden changes due to overall trends, and big players migration policies. Here we go deeper into showing the impact of big players infrastructure changes over the years, focusing in particular on the distance of servers from customers.



(a) CDF of RTT in 2014 and 2017 for Facebook and Instagram.



(b) CDF of RTT in 2014 and 2017 for YouTube and Google.

Fig. 12: Distribution of Round Trip Times.

A. *The birth of the sub-millisecond Internet*

CDNs were born in the '90s to reduce both the load on centralized servers and the delay to access the content. Shared and private CDNs make it possible to scale Internet content distribution, allowing users to fetch content from nearby servers. Since delay is a main parameter affecting users' experience, we focus on how it changed over years.

We first consider the Round Trip Time (RTT) as a performance index. Remind that probes measure RTT by matching client TCP segments with corresponding server TCP ACKs. We focus on the RTT from the probe to the server, thus excluding the access network delay. For all TCP connections to a given service in a long interval, we extract the minimum per-flow RTT, and compute the distribution of the number of flows according to RTT. We focus on the body of the distribution of minimum RTT per-flow, ignoring tails of the distribution, which may be caused by queuing and processing delays.

Figure 12 shows results, comparing measurements from April 2014 and April 2017. We focus on Facebook's and Google's services, since they are known for optimizing content delivery. Consider Instagram traffic (red curves) on Figure 12a. Dashed line refers to 2014, when there were already CDN nodes at 3 ms RTT from the ISP PoP. However, they served only 10% of the flows. Other traffic was served by CDN nodes further away, with RTT of 10, 20 and 30 ms.[7] About 7% of the flows was served by servers with RTT higher than 100 ms – a sign of intercontinental paths. Facebook caches (blue curves) follow a similar placement, with different share of traffic served by the caches.

[7]Fraction oscillates during the days. These figures refer to statistics collected on the whole month.

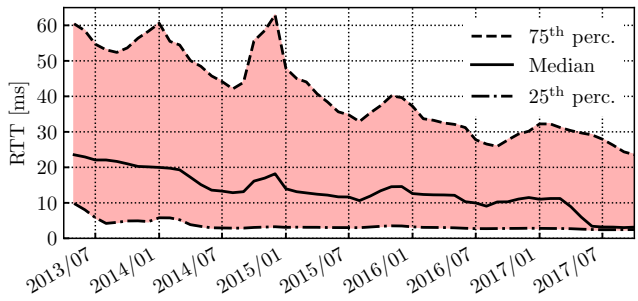


Fig. 13: Median, 15th and 75th percentile of bytes served from servers at a given RTT.

Moving to 2017 (solid lines), results clearly show that many more requests are now served by close servers, with 80 % of both Instagram and Facebook traffic served by 3 ms far CDN nodes. As we will see later, this change is due to two factors: (i) Facebook deployed its own CDN; and (ii) Instagram infrastructure has been integrated into Facebook.

Figure 12b depicts the RTT CDF for Google web search and YouTube streaming servers. In 2014, 80 % of YouTube traffic (dashed blue curve) was already served by nodes that were just 3 ms far away from the ISP PoP. In 2017, the already marginal RTT figure decreased more – with the YouTube video cache now breaking the sub millisecond RTT from the ISP PoP. That is, YouTube now directly places video servers inside the PoP, at the first level of aggregation, going further towards a very distributed and pervasive infrastructure. Interestingly, Google search servers (red curves) have not yet reached such a fine grained penetration, which is likely due its more diverse traffic when compared to YouTube video.

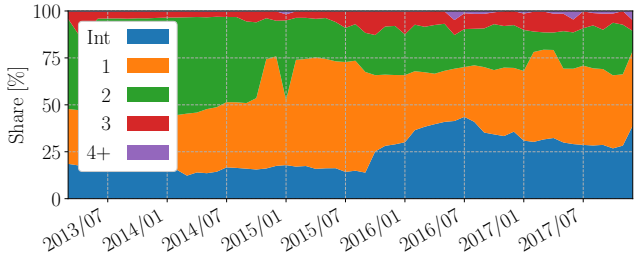
We have confirmed these findings by directly contacting the ISP staff, who reported the deployment of third-party CDN nodes at the ISP first aggregation points.

B. *How far is my content?*

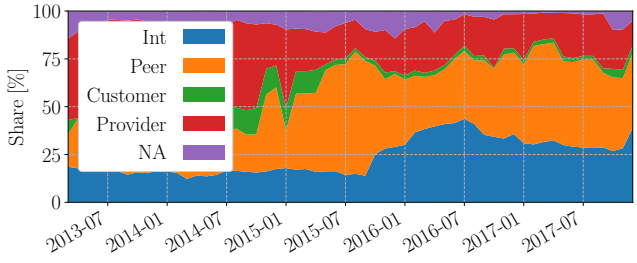
To generalize the above results, we study how the RTT distribution has changed over the years and the paths taken by packets to reach the ISP network.

1) RTT distance: We evaluate the RTT of all TCP flows (considering all customers and services), calculating the daily distribution of the minimum RTT over all flows. Results are in Figure 13, which shows the temporal evolution of the median, 15th and 75th percentiles of the daily distributions.

In the left hand side of the figure we can see that 75% of connections were served from servers at most 60 ms away in 2013. The median RTT in 2013 was around 20 ms. A clear decreasing trend is then observed. The median and the 15th percentile are already under 1 ms in 2017, thanks to the deployment of caches of popular services in the ISP PoPs. Notice the clear decrease in median RTT, which coincides with the deployment of the YouTube caches in this PoP. The tail of the distribution is also decreasing. In 2017, 75% of the flows go to servers closer than 25 ms RTT from customers.



(a) AS-level path lengths of incoming traffic reaching the ISP.



(b) Incoming traffic: "Internal" or reaching the ISP trough "Customer", "Peer" or "Provider" ASes. "NA" means that the AS relation cannot be determined.

Fig. 14: Breakdown of incoming traffic at the AS-level.

We have also evaluated RTT variability, i.e., the RTT span defined as the difference between the maximum and minimum RTT per flow. Our results confirm findings of [24], showing a decreasing trend from 2013 to 2015. Results are reported in Appendix B.

2) AS path distance: We analyze the paths of packets in terms of traversed ASes, computing the daily distribution of incoming bytes per AS path length. Figure 14a depicts the share of bytes served within the ISP (i.e., internal), networks directly connected to the ISP (AS path length equals to 1) and other networks (path length greater than 1).

We observe short-term fluctuations due to measurement artifacts, such as temporary changes in the paths reported by the CAIDA's Relations dataset. Despite these fluctuations, the AS path length shows a clear decreasing trend over time. In 2013 around 50% of the traffic arrived from internal servers (e.g., CDN nodes) or from ASes directly connected to the ISP. Less than 1% of the traffic arrived from ASes more than 3 hops away. The percentage of internal traffic and traffic from directly connected ASes has grown to 67% in 2017, while the share of traffic from distant ASes remained negligible.

To complete the analysis, we compute the relations of ASes delivering traffic to this ISP. Figure 14b shows how much traffic enters the ISP from *customer*, *peer*, or *provider* ASes. Notice that the share of traffic for which we cannot classify the ingress AS (*NA*) is generally less than 10%. The figure confirms the increase of internal traffic served by servers directly hosted in the ISP network. More interestingly, it shows an increase in traffic coming from peer ASes. Traffic from provider ASes has decreased, representing less than 20% of the incoming traffic for this ISP in 2017.

We confirmed with the ISP that the shifts between "peering" and "internal" were mainly due to YouTube service. Moreover, in November 2017, 90% of Netflix traffic shifted from peering to internal due to the deployment of Netflix caches in ISP premises too.

On the one hand, the proliferation of edge caches and the short delay of FTTH access networks are leading to the sub-millisecond Internet [42], This necessary to serve particular content, such as video [8] or advertisements [2]. On the other hand, this poses new burdens on the ISPs, which have to host (and in some cases manage) infrastructure of different content and CDN providers in their networks.

C. The Internet of few giants

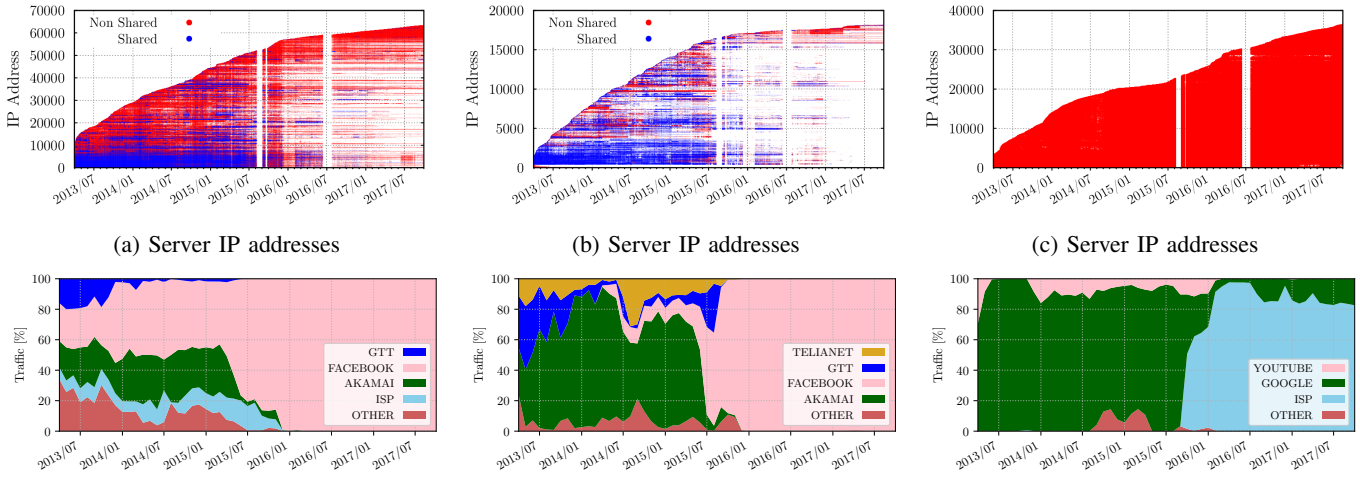
We finally analyze the infrastructure of popular services. Figure 15 depicts the evolution of the infrastructure of Facebook (left), Instagram (center), and YouTube (right) as seen from the ISP. Top plots show the server IP addresses active in each day for the considered service. The y-axis represents a single server IP address sorted in order of appearance. A red dot is present if the IP address was used only for traffic of the considered service in that day. A blue dot is present if the IP address served also content for other services. Finally, no dot is present if the IP address is not contacted in that day.

In all cases, we see new IP addresses appearing over time, counting several tens of thousands unique IP addresses. Compare Facebook and Instagram in Figure 15a and Figure 15b, respectively. During 2013 and 2014, a good fraction of addresses were shared with other services. During the second half of 2015, we notice major changes, with (i) a decrease in the number of contacted IP addresses, and (ii) an isolation of IP addresses, which are no longer shared with other services. The total number of IP addresses daily used by Facebook dropped from 3 800 to less than 1 000, out of which 700 are still shared. Since July 2016, shared IP addresses dropped further.

To better understand these changes, we analyze the ASes hosting the external IP addresses.[8] Bottom plots in Figure 15 show the daily traffic breakdown per AS for the services. Figure 15d and Figure 15e show a clear migration from generic CDNs to the Facebook private CDN. For Facebook, the migration started before 2013 and was completed in 2015. For Instagram, the integration with Facebook started in 2014 and was completed in 2015. This migration has two major effects: (i) IP addresses are dedicated to either Facebook or Instagram; (ii) the number of IP addresses contacted per day decreased. Contrasting these figures with Figure 12a, we notice that this change also benefited the RTT, which was reduced significantly.

Similarly, we depict the YouTube infrastructure evolution. From Figure 15c, it is already possible to see how it is different from the previous two cases. Indeed, YouTube always used a dedicated infrastructure. Its infrastructure keeps growing, with 40 000 IP addresses contacted daily in 2017.

[8]We use the RIB for each month from a major vantage point in the RouteViews to map IP addresses to ASNs



(a) Server IP addresses (b) Server IP addresses (c) Server IP addresses

(d) Destination Autonomous System (e) Destination Autonomous System (f) Destination Autonomous System

Fig. 15: Facebook (left), Instagram (center) and YouTube (right) infrastructure evolution over time.

These results confirm the trend towards a consolidation of large services, which deploy their own infrastructure in a more and more capillary way.

VII. Related Work and Discussion

Our work contributes with a new longitudinal dataset of traffic logs, highlighting major trends in Internet usage during a recent 5-year period. We are not the first to perform such an analysis, though. Here we discuss how trends seen in our data compare to previous measurements.

A. Internet usage evolution over the years

A number of measurement campaigns have been performed throughout the years [6], [7], [18], [19], [20], [22], [31]. While each study has its peculiarities, our work shares general goals and parts of the methodology with these previous efforts.

For example, authors of [19], [31] monitored traffic close to end users in the search for trends in traffic consumption. In [18] authors present a longitudinal study of Internet traffic covering the period of 1998–2003. Authors of [7] evaluated 7 years of MAWI traces, summarizing the evolution of Internet traffic in Japan. In [6], authors evaluated 23 months of data collected from 53 thousand broadband installations, highlighting the relation between capacity and demand.

Compared to previous works, we reassess trends and confirm most of their findings, while also highlighting new facts. Among key insights we can mention:

- The slow increase on average traffic per user [12]: We not only confirm it but also identify a relevant reduction in the tail of the distribution of upload traffic per broadband user;
- The latency reduction due to closer CDN nodes [24]: We identify interesting extreme cases, reporting the effect of providers deploying caches directly in ISP PoPs;
- We find that the decline of P2P traffic [21], [31] has reached a plateau with a small but persistent volume of traffic;
- We confirm the predominance of video traffic [1], [16], [32]. With the arrival of Ultra HD, we highlight large differences on video consumption among Fiber and ADSL customers;
- We could not find a clear relation between the capacity offered to customers and their demands, as in [6]. However, for customers relying on particular services (like Netflix, see above) these conclusions seem to hold;
- We quantify for the first time the impact of new social networks, such as Instagram and Snapchat, on ISP networks, finding that in some cases their traffic is comparable to video-on-demand services;
- We witness the fast increase in HTTPS deployment [17] and document experiments with new protocols such as HTTP/2, SPDY, QUIC and FB-ZERO. We observe sudden changes in traffic mix due to bugs and private tests by large companies [26], [28], [39]. We find that these new and experimental protocols account for almost 40% of the web traffic in 2017;
- We confirm and quantify the concentration of Internet traffic around a few big players [27], [32]. Moreover, we observe an increase on the share of traffic served within the ISP or from AS peers, in part due to the deployment of caches and CDN nodes from such players inside ISP premises.

B. Other measurement perspectives

Several works perform an analysis similar to ours, but taking different points of view. These works include those measuring from different locations [7], in mobile networks [29], [32], [40], in core networks [27], [38], among others.

Liu *et al.* [29] designed a large-scale measurement infrastructure and deployed it in the core network of a cellular operator. Muhammad *et al.* [40] analyze traffic trace from a tier-1 cellular operator. Authors of [32] analyze traffic logs collected in a 3G/4G network deployed in a major European country. This work is particularly interesting since data has been collected geographically close to our monitoring points and in a period overlapping with our measurements.

Our work is *complementary* to these efforts. Indeed, we claim that the type of customers under study still impact traffic profiles strongly. For example, authors of [40] show how machine-to-machine traffic in mobile networks is different from human-generated traffic. Marquez et al. [32] show the mix of applications dominating download traffic in a 3G/4G network, which is largely dissimilar from what we observe in broadband networks. Nevertheless, our measurements show that the use of mobile devices connected to WiFi at home is reducing such differences, e.g., with mobile apps such as Instagram and Facebook among the top applications in both scenarios.

Given the difficulties of obtaining relevant measurements, different efforts are needed to have a comprehensive picture of the Internet. We believe the figures we presented in this paper are vital to researchers, ISPs and even web service providers to better understand the liveliness of the Internet, which continuously evolves, mixing slow and unpredictable changes. To broaden the impact of our contributions, we make anonymized data available to the community at https://smartdata.polito.it/five-years-at-the-edge/.

Acknowledgments

The research leading to these results has been funded by both the Vienna Science and Technology Fund (WWTF) through project ICT15-129 (BigDAMA) and the SmartData@PoliTO center for Big Data technologies. We would like to thank also the technical teams from the ISP providing data for this research for their continuous collaboration and support.

References

[1] V. K. Adhikari, S. Jain, and Z.-L. Zhang. YouTube Traffic Dynamics and Its Interplay with a Tier-1 ISP: An ISP Perspective. In *Proceedings of the ACM SIGCOMM Internet Measurement Conference*, IMC'10, pages 431–443, 2010.

[2] P. Bangera and S. Gorinsky. Ads versus regular contents: Dissecting the web hosting ecosystem. In *2017 IFIP Networking Conference (IFIP Networking) and Workshops*, pages 1–9. IEEE, 2017.

[3] P. Barford and M. Crovella. Measuring Web Performance in the Wide Area. *SIGMETRICS Perform. Eval. Rev.*, 27(2):37–48, 1999.

[4] M. Belshe, R. Peon, and M. Thomson. Hypertext Transfer Protocol Version 2 (HTTP/2). Technical Report 7540, RFC Editor, 2015.

[5] I. Bermudez, M. Mellia, M. M. Munafò, R. Keralapura, and A. Nucci. DNS to the Rescue: Discerning Content and Services in a Tangled Web. In *Proceedings of the ACM SIGCOMM Internet Measurement Conference*, IMC'12, pages 413–426, 2012.

[6] Z. S. Bischof, F. E. Bustamante, and R. Stanojevic. Need, Want, Can Afford: Broadband Markets and the Behavior of Users. In *Proceedings of the ACM SIGCOMM Internet Measurement Conference*, IMC'14, pages 73–86, 2014.

[7] P. Borgnat, G. Dewaele, K. Fukuda, P. Abry, and K. Cho. Seven Years and One Day: Sketching the Evolution of Internet Traffic. In *IEEE INFOCOM 2009*, INFOCOM'09, pages 711–719, 2009.

[8] T. Böttger, F. Cuadrado, G. Tyson, I. Castro, and S. Uhlig. Open connect everywhere: A glimpse at the internet ecosystem through the lens of the netflix cdn. *ACM SIGCOMM Computer Communication Review*, 48(1):28–34, 2018.

[9] CAIDA. As relationships, customer cones, and validation.

[10] CAIDA. Routeviews prefix to AS mappings dataset (pfx2as) for IPv4 and IPv6.

[11] I. Castro, J. C. Cardona, S. Gorinsky, and P. Francois. Remote peering: More peering without internet flattening. In *Proceedings of the 10th ACM International on Conference on Emerging Networking Experiments and Technologies*, CoNEXT '14, pages 185–198, New York, NY, USA, 2014. ACM.

[12] K. Cho, K. Fukuda, H. Esaki, and A. Kato. Observing Slow Crustal Movement in Residential User Traffic. In *Proceedings of the 2008 ACM CoNEXT Conference*, CoNEXT'08, pages 1–12, 2008.

[13] K. C. Claffy. Measuring the Internet. *IEEE Internet Computing*, 4(1):73–75, 2000.

[14] J. Dean and S. Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. *Commun. ACM*, 51(1):107–113, 2008.

[15] A. Dhamdhere and C. Dovrolis. Twelve Years in the Evolution of the Internet Ecosystem. *IEEE/ACM Trans. Netw.*, 19(5):1420–1433, 2011.

[16] J. Erman, A. Gerber, K. Ramadrishnan, S. Sen, and O. Spatscheck. Over the Top Video: The Gorilla in Cellular Networks. In *Proceedings of the ACM SIGCOMM Internet Measurement Conference*, IMC'11, pages 127–136, 2011.

[17] A. P. Felt, R. Barnes, A. King, C. Palmer, C. Bentzel, and P. Tabriz. Measuring HTTPS Adoption on the Web. In *Proceedings of the 26th USENIX Security Symposium*, USENIX Security'17, pages 1323–1338, 2017.

[18] M. Fomenkov, K. Keys, D. Moore, and K. C. Claffy. Longitudinal Study of Internet Traffic in 1998-2003. In *Proceedings of the Winter International Synposium on Information and Communication Technologies*, WISICT'04, pages 1–6, 2004.

[19] C. Fraleigh, S. Moon, B. Lyles, C. Cotton, M. Khan, D. Moll, R. Rockell, T. Seely, and C. Diot. Packet-level Traffic Measurements from the Sprint IP Backbone. *Netwrk. Mag. of Global Internetwkg.*, 17(6):6–16, 2003.

[20] J. L. García-Dorado, A. Finamore, M. Mellia, M. Meo, and M. M. Munafò. Characterization of ISP Traffic: Trends, User Habits, and Access Technology Impact. *IEEE Trans. Netw. Service Manag.*, 9(2):142–155, 2012.

[21] A. Gavalda-Miralles, D. R. Choffnes, J. S. Otto, M. A. Sanchez, F. E. Bustamante, L. Amaral, J. Duch, and R. Guimera. Impact of Heterogeneity and Socioeconomic Factors on Individual Behavior in Decentralized Sharing Ecosystems. *Proceedings of the National Academy of Sciences*, 111(43):15322–15327, 2014.

[22] S. Gebert, R. Pries, D. Schlosser, and K. Heck. Internet Access Traffic Measurement and Analysis. In *Proceedings of the 4th International Conference on Traffic Monitoring and Analysis*, TMA'12, pages 29–42, 2012.

[23] R. Hofstede, P. Čeleda, B. Trammell, I. Drago, R. Sadre, A. Sperotto, and A. Pras. Flow Monitoring Explained: From Packet Capture to Data Analysis with NetFlow and IPFIX. *Commun. Surveys Tuts.*, 16(4):2037–2064, 2014.

[24] T. Høiland-Jørgensen, B. Ahlgren, P. Hurtig, and A. Brunstrom. Measuring latency variation in the internet. In *Proceedings of the 12th International on Conference on emerging Networking EXperiments and Technologies*, pages 473–480. ACM, 2016.

[25] Intel. DPDK - Data Plane Development Kit, 2011.

[26] A. M. Kakhki, S. Jero, D. Choffnes, C. Nita-Rotaru, and A. Mislove. Taking a Long Look at QUIC: An Approach for Rigorous Evaluation of Rapidly Evolving Transport Protocols. In *Proceedings of the ACM SIGCOMM Internet Measurement Conference*, IMC'17, pages 290–303, 2017.

[27] C. Labovitz, S. Iekel-Johnson, D. McPherson, J. Oberheide, and F. Jahanian. Internet Inter-Domain Traffic. In *Proceedings of the ACM Conference on Data Communication*, SIGCOMM'10, pages 75–86, 2010.

[28] A. Langley, J. Iyengar, J. Bailey, J. Dorfman, J. Roskind, J. Kulik, P. Westin, R. Tenneti, R. Shade, R. Hamilton, V. Vasiliev, A. Riddoch, W.-T. Chang, Z. Shi, A. Wilk, A. Vicente, C. Krasic, D. Zhang, F. Yang, F. Kouranov, and I. Swett. The QUIC Transport Protocol: Design and Internet-Scale Deployment. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication*, SIGCOMM'17, pages 183–196, 2017.

[29] J. Liu, F. Liu, and N. Ansari. Monitoring and Analyzing Big Traffic Data of a Large-Scale Cellular Network with Hadoop. *Netwrk. Mag. of Global Internetwkg.*, 28(4):32–39, 2014.

[30] Lixin Gao. On inferring autonomous system relationships in the internet. *IEEE/ACM Transactions on Networking*, 9(6):733–745, Dec 2001.

[31] G. Maier, A. Feldmann, V. Paxson, and M. Allman. On Dominant Characteristics of Residential Broadband Internet Traffic. In *Proceedings of the ACM SIGCOMM Internet Measurement Conference*, IMC'09, pages 90–102, 2009.

[32] C. Marquez, M. Gramaglia, M. Fiore, A. Banchs, C. Ziemlicki, and Z. Smoreda. Not all apps are created equal: Analysis of spatiotemporal heterogeneity in nationwide mobile service usage. In *Proceedings of the 13th International Conference on emerging Networking EXperiments and Technologies*, pages 180–186. ACM, 2017.

[33] M. Mellia, M. Meo, L. Muscariello, and D. Rossi. Passive Identification and Analysis of TCP Anomalies. In *Proceedings of the IEEE International Conference on Communications*, ICC'06, pages 723–728, 2006.

[34] D. Moore, C. Shannon, D. J. Brown, G. M. Voelker, and S. Savage. Inferring Internet Denial-of-service Activity. *ACM Trans. Comput. Syst.*, 24(2):115–139, 2006.

[35] V. Moreno, J. Ramos, P. M. S. del Río, J. L. García-Dorado, F. J. Gomez-Arribas, and J. Aracil. Commodity Packet Capture Engines: Tutorial, Cookbook and Applicability. *Commun. Surveys Tuts.*, 17(3):1364–1390, 2015.

[36] D. Naylor, A. Finamore, I. Leontiadis, Y. Grunenberger, M. Mellia, M. Munafò, K. Papagiannaki, and P. Steenkiste. The Cost of the "S" in HTTPS. In *Proceedings of the 10th ACM International Conference on Emerging Networking Experiments and Technologies*, CoNEXT'14, pages 133–140, 2014.

[37] C. Orsini, A. King, D. Giordano, V. Giotsas, and A. Dainotti. BGPStream: A Software Framework for Live and Historical BGP Data Analysis. In *Proceedings of the 2016 ACM Internet Measurement Conference*, IMC'16, pages 429–444, 2016.

[38] P. Richter, N. Chatzis, G. Smaragdakis, A. Feldmann, and W. Willinger. Distilling the Internet's Application Mix from Packet-Sampled Traffic. In *Proceedings of the 16th International Conference on Passive and Active Measurement*, PAM'15, pages 179–192, 2015.

[39] J. Rüth, I. Poese, C. Dietzel, and O. Hohlfeld. A First Look at QUIC in the Wild. In *Proceedings of the 19th International Conference on Passive and Active Measurement*, PAM'18, pages 255–268, 2018.

[40] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang. A First Look at Cellular Machine-to-machine Traffic: Large Scale Measurement and Characterization. In *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS'12, pages 65–76, 2012.

[41] Y. Shavitt and E. Shir. DIMES: Let the Internet Measure Itself. *SIGCOMM Comput. Commun. Rev.*, 35(5):71–74, 2005.

[42] A. Singla, B. Chandrasekaran, P. B. Godfrey, and B. Maggs. The Internet at the Speed of Light. In *Proceedings of the 13th ACM Workshop on Hot Topics in Networks*, HotNets'14, pages 1–7, 2014.

[43] M. Trevisan, I. Drago, M. Mellia, and M. M. Munafò. Towards Web Service Classification using Addresses and DNS. In *Proceedings of the 7th International Workshop on Traffic Analysis and Characterization*, TRAC'16, pages 38–43, 2016.

[44] M. Trevisan, A. Finamore, M. Mellia, M. Munafo, and D. Rossi. Traffic Analysis with Off-the-Shelf Hardware: Challenges and Lessons Learned. *IEEE Commun. Mag.*, 55(3):163–169, 2017.

[45] M. Trevisan, D. Giordano, I. Drago, M. Mellia, and M. Munafo. Five Years at the Edge: Watching Internet from the ISP Network. In *Proceedings of the 14th International Conference on Emerging Networking EXperiments and Technologies*, CoNEXT'18, pages 1–12, 2018.

[46] L. Vassio, I. Drago, M. Mellia, Z. B. Houidi, and M. L. Lamali. You, the web, and your device: longitudinal characterization of browsing habits. *ACM Transactions on the Web (TWEB)*, 12(4):24, 2018.

[47] C. Williamson. Internet Traffic Measurement. *IEEE Internet Computing*, 5(6):70–74, 2001.

[48] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica. Spark: Cluster Computing with Working Sets. In *Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing*, HotCloud'10, pages 10–10, 2010.

BIOGRAPHIES

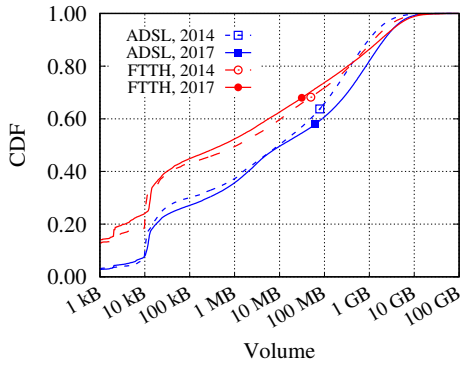
Martino Trevisan received his PhD in 2019 from Politecnico di Torino, Italy. He is currently a post-doctoral researcher at the SmartData@Polito center in the same university. He has been collaborating in both Industry and European projects and spent six months in Telecom ParisTech, France working on High-Speed Traffic Monitoring during his M.Sc. He visited twice Cisco labs in San Jose in summer 2016 and 2017, as well as AT&T labs during fall 2018. His research interests are Big Data and Machine Learning applied to Network Measurements and Traffic Monitoring.

Danilo Giordano Ph.D., is an Assistant Professor at Politecnico di Torino. He is member at the SmartData@Polito lab, working at the application of big data analysis of networks and applied to smart cities and predictive maintenance problems. During 2014, he was a research intern at Narus Inc., applying big data analytics to network anomaly detection. In 2016, he has been with CAIDA at UCSD working on network traffic characterization, implementing a large scale system to collect and analyze BGP data. For this topic, he won the Applied Networking Research Prize by IETF. His broaden interests are in the fields of data analysis, big data, and networking.

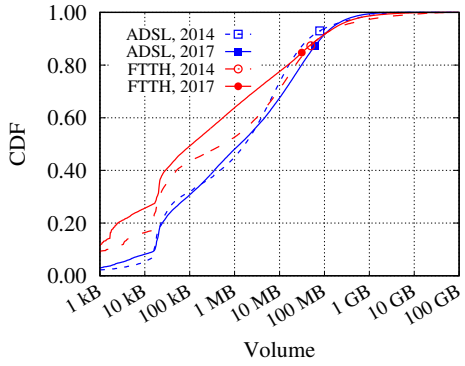
Idilio Drago is an Assistant Professor (RTDa) at the Politecnico di Torino, Italy, in the Department of Electronics and Telecommunications. His research interests include Internet measurements, Big Data analysis, and network security. Drago has a PhD in computer science from the University of Twente. He was awarded an Applied Networking Research Prize in 2013 by the IETF/IRTF for his work on cloud storage traffic analysis.

Maurizio Matteo Munafò is Assistant Professor at the Department of Electronics and Telecommunications of Politecnico di Torino. He holds a Dr.Ing. degree in Electronic Engineering since 1991 and a Ph.D. in Telecommunications Engineering since 1994, both from Politecnico di Torino. He has co-authored about 80 journal and conference papers in the area of communication networks and systems. His current research interests are in simulation and performance analysis of communication systems and traffic modeling, measurement, and classification.

Marco Mellia (S'08), Ph.D., research interests are in the in the area of traffic monitoring and analysis, in cyber monitoring, and Big Data analytics. Marco Mellia has co-authored over 250 papers published in international journals and presented in leading international conferences. He won the IRTF ANR Prize at IETF-88, and best paper award at IEEE P2P'12, ACM CoNEXT'13, IEEE ICDCS'15. He is part of the editorial board of ACM/IEEE Transactions on Networking, IEEE Transactions on Network and Service Management, and ACM Computer Communication Review. He holds a position as Full Professor at Politecnico di Torino, Italy.



(a) Download



(b) Upload

Fig. 16: ECDFs of downloaded and uploaded volumes per customer per day

APPENDIX A
ACTIVE CUSTOMERS THRESHOLD SETTING

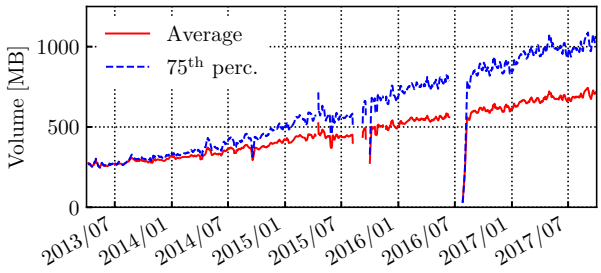
We discuss and detail some methodological choices we made for our analysis. Consider the active customer definition. We consider a customers as active during a given day if it generated at least 10 flows, *and* downloaded at least 15 kB *and* uploaded at least 5 kB. We set these threshold based both on domain knowledge, and on the distributions of those metrics. For instance, a complete TLS handshake typically takes 5 to 10 kB to download certificates and negotiate keys. Hence, an active user surfing the web should produce more data.

Figure 16 details the ECDFs of downloaded (16a) and uploaded (16b) bytes per customer per day. The Download figure shows a clear knee at about 15 kB which allows us to filters out a large number of households with very low downloaded volume. We take a more conservative threshold on the upload data (bottom plot) since we expect customers to upload less. Cross checking the filters, we confirm that most of the customers removed by the download threshold are the same uploading negligible volume of data. After removing these inactive customers, we obtain Figure2 in the paper.

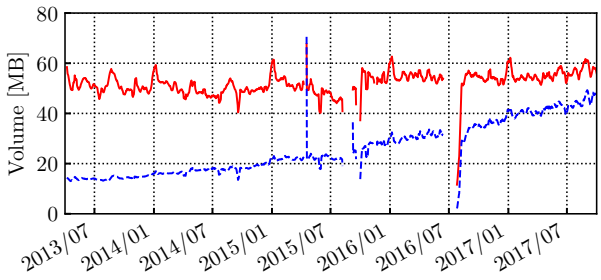
APPENDIX B
SUPPLEMENTARY RESULTS

In this appendix we report supplementary results left out for the sake of brevity.

A. Subscriber's traffic consumption in detail



(a) Download



(b) Upload

Fig. 17: Average and 75th percentile of per-subscription daily traffic (ADSL).

Here we briefly offer more details about the distribution of daily per-subscriber traffic. As already stated in Section III-A, the average per-subscriber traffic volume increased for download while remained almost constant for upload (for both ADSL and FTTH subscribers). However, analyzing more in depth traffic distributions, further considerations hold. Figure 2 already suggests changes in the tails of the distributions, i.e., heavy-consuming uploaders are less frequent. We further quantify this phenomenon in Figure 17, where we report both average (red curve) and the 75th percentile (blue dashed curve) of per-subscription daily traffic. The figure only depicts ADSL users, as a similar picture emerges for FTTH. Considering download (Figure 17a), on 2013 average and 75th percentile assumed similar values, in the range of [250-270] MB/day. Similar values happen due to the heavy tail of the distribution, meaning that heavy-consuming subscribers are sizable. As years pass, the 75th percentile increases at a higher pace than the average (red solid line), pinpointing to fewer heavy-consuming subscribers. Similar considerations hold for upload (Figure 17b). On 2013 average was no less than three times the 75th percentile. While the average shows a flat trend over the years, the 75th percentile constantly increases, leading to similar conclusions.

B. Traffic volume of popular services

Figure 18 complements Figure 5, and depicts a similar picture for the percentage of downloaded bytes for each service. The multi-color palette is set to 10% to improve the visualization. We can observe how services have changed their con-

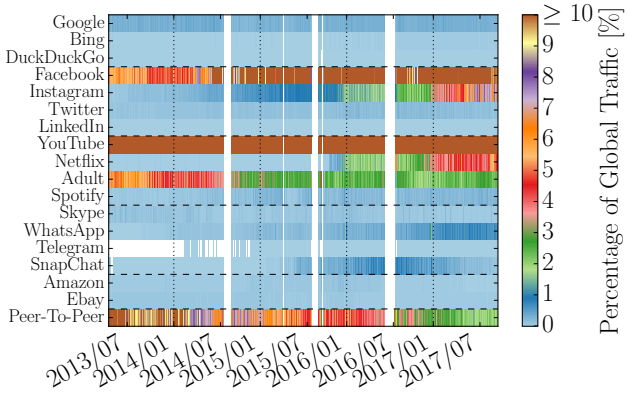
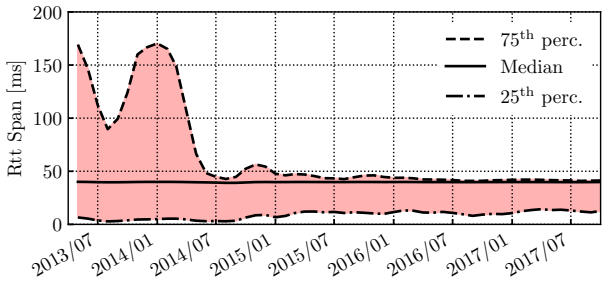
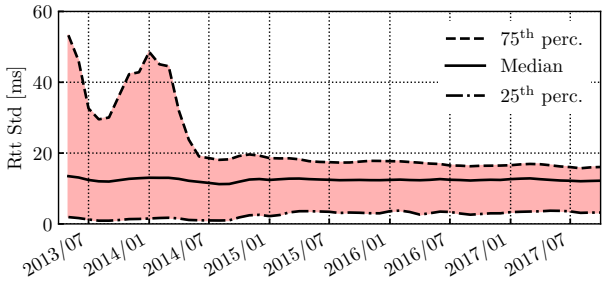


Fig. 18: Percentage of downloaded bytes for selected services over time.



(a) RTT Span distribution



(b) RTT Standard Deviation distribution

Fig. 19: Median, 15th and 75th percentile of the Span and Standard Deviation of RTT.

C. RTT Span

Figure 19 shows the trend of RTT variability over time, detailing the 25th, 50th and 75th percentiles of the RTT span (Figure 19a) and RTT standard deviation (Figure 19b). Recall that the RTT Span is the difference between maximum and minimum RTT observed within a TCP connection [24]. As stated in Section VI, we observe a consistent reduction of both metrics during 2014, followed by a more flat picture.

tributions to the traffic mix during the monitored period. Notice, for instance, how services such as Facebook, Instagram, WhatsApp and Netflix have increased traffic share throughout the years. Others, such as SnapChat have gained momentum only during a limited period.

D. Remote Peering

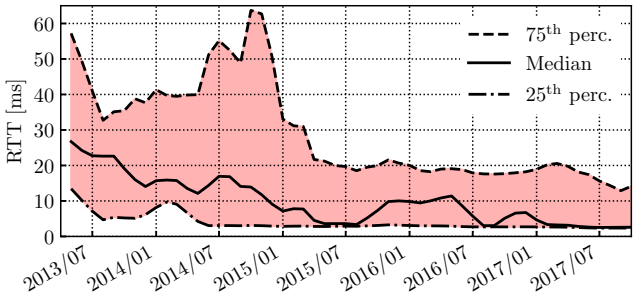


Fig. 20: Median, 15th and 75th percentile of RTT toward a Peering destination.

To reinforce results in [11] we analyze the RTT for flows originated by servers located in peering ASes according to the CAIDA Relationship Dataset [9]. We aim at quantifying the impact of remote peering – i.e., the practice of establishing remote or indirect peering connections either over a "long cable" (owned or leased) or over resellers. Figure 20 reports median and percentiles of the RTT for flows towards peering ASes. Although we do not provide a per-peer in-depth analysis, this result again shows a reduction of RTT percentiles over the years, which lets us conjecture that remote peering is being reduced, if present.