Mind The Gap Between HTTP and HTTPS in Mobile Networks

Alessandro Finamore, Matteo Varvello, Kostantina Papagiannaki

Telefonica Research
{name.surname}@telefonica.com

Abstract. Fueled by a plethora of applications and Internet services, mobile data consumption is on the rise. Over the years, mobile operators deployed webproxies to optimize HTTP content delivery. Webproxies also produce HTTP-logs which are a fundamental data source to understand network/services performance and user behavior. The recent surge of HTTPS is progressively reducing such wealth of information, to the point that it is unclear whether HTTP-logs are still representative of the overall traffic. Unfortunately, HTTPS monitoring is challenging and adds some extra cost which refrains operators from "turning on the switch". In this work, we study the "gap" between HTTP and HTTPS both quantifying their intrinsic traffic characteristics, and investigating the usability of the information that can be logged from their transactions. We leverage a 24-hours dataset collected from a webproxy operated by a European mobile carrier with more than 10M subscribers. Our quantification of this gap suggests that its importance is strictly related to the target analysis.

1 Introduction

Mobile operators are facing an explosion of demand for data access services. Recent estimates forecast an eight-fold increase of demand between 2015 and 2020, a rate three times higher than for fixed access networks [15]. This explosion is driven both by the constant evolution of the mobile apps and Internet services ecosystem, and the roll out of 4G technologies.

In this dynamic and demanding scenario, traffic monitoring is paramount. Accurate understanding of both user behavior and service quality are key to drive network investments. To study data services, mobile operators rely on *Usage Data Records (UDRs)* and *HTTP-logs*. UDRs aggregate users data activity over periods of time lasting from minutes up to multiple hours. They are collected for billing purposes and do not detail the apps/services used [8, 9].

Differently from UDRs, HTTP-logs contain detailed information on individual HTTP transactions. They are usually collected by *webproxies*, middle-boxes that aim at optimizing HTTP delivery through in-network caching and content modification (e.g., image resolution reduction) [4, 13]. HTTP-logs have been extensively used both by operators and academia to characterize mobile network traffic [2, 3, 6, 11, 12, 16].

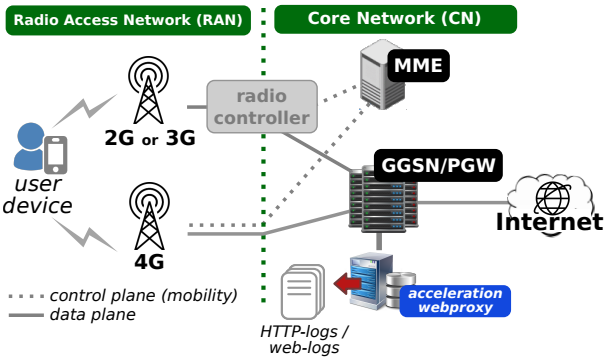

Fig. 1. Sketch of a mobile network architecture where a web acceleration proxy is deployed.

With the rise of HTTPS, this scenario is however rapidly changing. For instance, between June 2015 and June 2016,[1] Google reported a +13% increase of requests served over HTTPS. Sandvine also reports that more than 60% of mobile traffic worldwide is currently encrypted [10]. While this calls for instrumenting webproxies to also log HTTPS transactions, it is unclear whether the additional cost is justified. In fact, HTTPS exposes little information about the service and content users access. In addition, network performance indexes (e.g., throughput and latency) can only be computed on the whole TLS connection and not on individual transactions, as commonly done for HTTP.

In this work we present the first comparative study between HTTP and HTTPS traffic for mobile networks. Our goal is to quantify the "gap" between HTTP and HTTPS both in term of their macroscopic qualities and of their accuracy when singularly used to perform common analysis such as data consumption, user mobility, etc. The input of our study is a unique dataset spanning HTTP and HTTPS traffic, radio-layer information, and device information from a 10M-subscriber European mobile operator.

Our quantification of this gap suggests that its importance is strictly related to the target analysis. When focusing on volume, neither HTTP nor HTTPS alone are enough to characterize users activity. This is because of a combination of factors including type of device and usage pattern across time. Conversely, both traffic types are capable to capture human-driven behaviors like user mobility, which in turn drives analysis like traffic consumption in space and cell towers utilization.

2 Background

This section overviews the classic mobile network architecture while emphasizing the role of webproxies in it (Fig. 1). The Radio Access Network (RAN),

[1] `https://www.google.com/transparencyreport/https/?hl=en`

commonly called "last mile", is composed of thousands of elements such as cell sectors, towers, and radio controllers. The Core Network (CN) bridges the RAN with the Internet by mean of packet data gateways (GGSN and PGW) which allow mobile users to access data services. The Mobility Management Entity (MME) servers handle *network events* related to handovers, paging, and access control to radio channels, each carrying the device id and the sector from which the event was triggered. The MME is the *control plane* of a mobile network.

Fig. 1 also shows an *acceleration webproxy*; this is a transparent (or explicit) HTTP proxy that operators deploy to speed up content delivery at the RAN while reducing traffic volume at the CN. Common webproxy services are: i) content caching, ii) content compression (e.g., reducing image size/resolution or video format re-encoding), and iii) dynamic traffic policies enforcement (e.g., bandwidth throttling for users that reach their monthly data cap, protection from malware and third party tracking services). Webproxies log each HTTP transaction into HTTP-logs, but some vendors provide monitoring solutions that also log the remaining TCP activity [10]. We call such "extended" logs *web-logs*.

3 Dataset

We consider web-logs collected for 24 consecutive hours (April 27th, 2016) by the acceleration webproxy of a major European mobile operator serving more than 10M subscribers. The considered webproxy usually logs HTTP traffic only, but it can be sporadically instrumented to report on other TCP traffic like HTTPS.

We call *transaction* an entry in the web-logs. Each transaction contains at least the following fields: IPs/ports tuple (source and destination), timestamp, duration, user-id, and bytes delivered. Additional fields can be provided based on the transaction type. Specifically, an "HTTP transaction" corresponds to an HTTP request/response exchange for which the webproxy further logs HTTP meta-data such as hostname, URL, user-agent, and content-type. User privacy is guaranteed by hashing sensible information like user-id, requested URL, etc. For the remainder of the traffic, a transaction corresponds to a TCP connection. If the `ClientHello` message from a TLS handshake is detected, the webproxy also logs the Service Name Identification (SNI), when provided.

We combine the webproxy dataset with two additional data sources.

Radio-layers enrichment: We process MME network events (see Sec. 2) to create *mobility radio-layers*, *i.e.,* per user timelines detailing to which sectors each user's device connects to over time. It follows that given the tuple (user-id, timestamp, duration) of a web-log transaction we can identify the list of sectors the transaction relates to. This enables us to investigate how content is consumed by users while moving across the network (see Sec. 5) at a finer granularity with respect to the literature [12, 16].

TAC enrichment: The Type Allocation Code (TAC) database is an internal resource of the considered operator, and it is based on the GSMA TAC database,[2]

[2] `https://imeidb.gsma.com/imei/login.jsp`

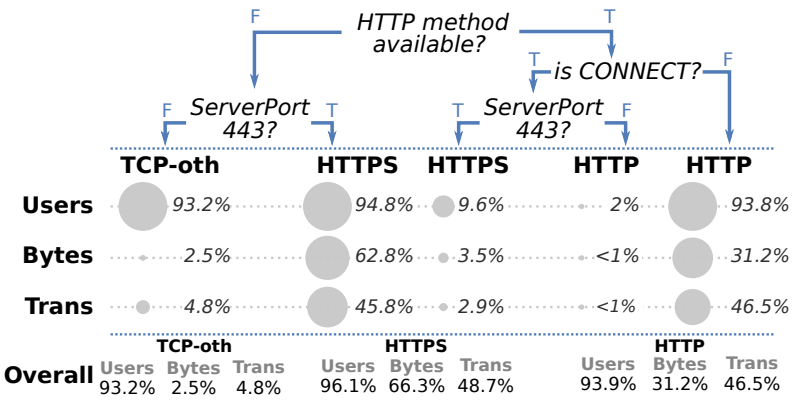

Fig. 2. Dataset overview.

i.e., the standardized allocation of TAC among vendors.[3] The TAC database is a static table mapping vendor and device model to a user-id. This mapping is more robust than the classic approach based on HTTP user-agent string, and it works also in presence of HTTPS.

3.1 Dataset Curation

Following the logic described in Fig. 2, we split web-log transactions into three classes: HTTP, HTTPS, and TCP-oth (*i.e.,* the TCP traffic that is neither HTTP nor HTTPS). Since the webproxy does not explicitly label web-log entries originated by TLS traffic, we identify HTTPS based on the destination port (443). The webproxy logs the `HTTP METHOD` for each HTTP transaction which eases the identification of HTTP traffic. In presence of a `CONNECT`, *i.e.,* for clients explicitly connecting through a proxy, we still use the destination port to distinguish between HTTP and HTTPS.

The middle of Fig. 2 reports the percentage of users, bytes, and transactions of each classification tree leaf, while aggregated statistics are reported at the bottom. Overall, HTTPS dominates the volume of bytes (66.3%) but we also find a non negligible 2.5% of TCP-oth volume. When we focus on transactions, we notice that they are equally distributed between HTTPS (48.7%) and HTTP (46.5%), which is counter-intuitive due to the bytes difference observed above. This is due to the presence of *persistent connections* that go undetected in HTTPS. We further analyze this issue in the following.

Content consumed in mobile networks is usually "small", e.g., the average object size is in the order of tens of kB [5, 16]. To reduce the TCP handshake overhead, HTTP 1.1 introduced the concept of persistent connections which allow devices to use a single TCP connection to send multiple requests. Such technique is common to both HTTP and HTTPS, but it is a hassle only when monitoring

[3] The TAC is part of the IMEI, *i.e.,* the unique identifier of a mobile device.

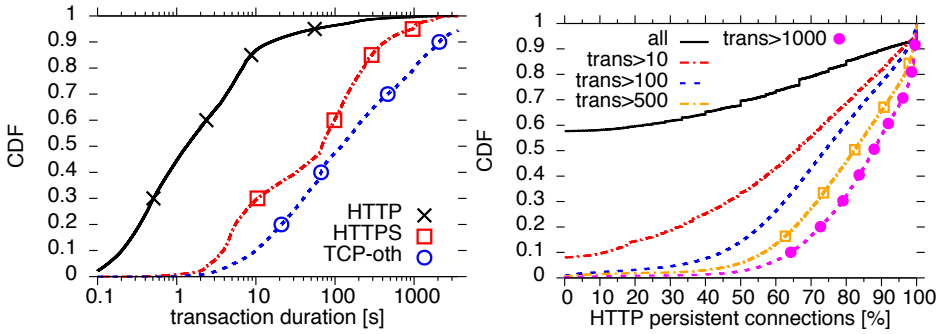

Fig. 3. Impact of persistent connections: transaction duration (left) and percentage of HTTP persistent connections (right).

HTTPS. In fact, encryption does not allow to identify request/response pairs over the same TLS connection, resulting in a coarser view over HTTPS traffic if compared to HTTP.

To visualize the impact of the latter limitation, Fig. 3 (left) shows the Cumulative Distribution Function (CDF) of the transaction duration for HTTP, HTTPS, and TCP-oth. If on the one hand the three traffic classes are subject to different dynamics due to how different services use them, on the other hand such huge differences hint to the presence of persistent connections.

We further corroborate on this by counting the percentage of TCP connections having more than one HTTP transaction for each user. Fig. 3 (right) shows the CDF of the fraction of persistent HTTP connections during one peek hour (results hold for different hours). The figure shows that the usage of persistent connections is indeed extremely common and proportional to user activity, e.g., 90% of the very active users (trans>1,000 in the plot) have more than 65% of their HTTP connections being persistent.

To the best of our knowledge, no previous study has quantified the adoption of persistent connections in the wild. Our results indicate that their high popularity can introduce substantial errors when comparing HTTP with HTTPS traffic. Accordingly, to enable a meaningful comparison among the considered traffic classes, we have opted for pre-processing HTTP traffic to aggregate different transactions belonging to the same individual TCP connections.

4 Overall Volumes

We start our analysis with a top-down characterization of how traffic volume (bytes) is split between traffic types.

Daily aggregate breakdown: Fig. 4 (left) shows the CDF of the percentage of HTTP, HTTPS, and TCP-oth volume, per user. As expected, HTTPS is the dominant traffic type: 50% of users have more than 77.6% of their volume carried

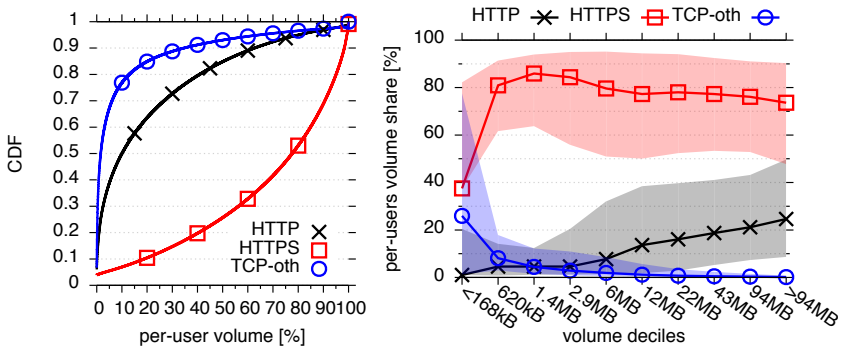

Fig. 4. Comparing traffic volume: on the left, overall percentage of per user HTTP, HTTPS, and TCP-oth; on the right, further breakdown with respect to absolute consumption (lines reflect the 50th percentile of each volume decile, while shaded areas indicate 25th-75th).

over encrypted connections. The figure also shows that TCP-oth volume is far from being negligible: 5.6% of users have more than 70% of TCP-oth traffic.

We further dig into the TCP-oth traffic using the destination port to classify the specific service being used. We find that 84% of volume is associated to email (e.g., 995/IMAP-SSL, 993/POP3-SSL, 110/POP3) and push notification services (5223 for Apple, 5228 for Android). We also find a few thousands users with "suspicious" behaviors: they contact 227k IP addresses using 49k ports (in a peak hour) and do not transfer any data on the opened TCP connection. For these 227k IP addresses, we further retrieve the Autonomous System Number (ASN) using Team Cymru[4] and its classification using PeeringDB[5] and CAIDA AS ranking [1]. Such analysis reveals that 97% of these IPs belong to fixed and mobile ISPs, and are not linked to classic services. We conjecture the presence of malware, of which we also find evidences,[6] or some form of P2P communication.

We further divide users into ten groups based on the deciles of the distribution of their volume consumption. For each group we then extract the 25th, 50th, and 75h percentiles of the share of HTTP, HTTPS, and TCP-oth. Fig 4 (right) reports the results (the x-axis details the used deciles). Beside noticing that HTTPS dominates indistinctly within each bin, we observe that TCP-oth shares are inversely proportional to the overall volume consumed, while the opposite is true for HTTP. Results reported in Fig. 4 hold also when considering the number of transactions (we avoid reporting them for brevity). Those differences are possibly due to the combination of apps/services used, but to the best of our knowledge, there are not robust techniques available to classify mobile traffic. Hence we leave a detailed characterization for the future.

[4] `http://www.team-cymru.org/IP-ASN-mapping.html#dns`
[5] `www.peeringdb.com`
[6] `http://bit.ly/1Uv9hNF`

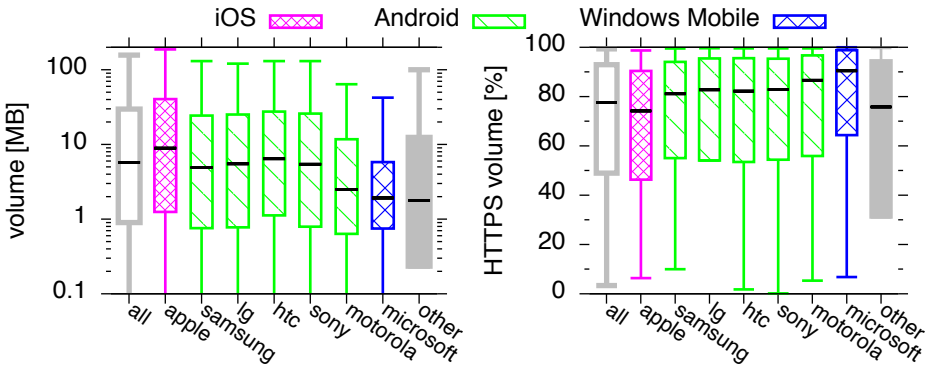

Fig. 5. Comparing OS per-device volume: daily aggregate (left) and related percentage of HTTPS (right).

Device type: We here investigate the relationship between device type and consumed volume. Fig. 5 (left) shows the boxplots (5th, 25th, 50th, 75th, 95th percentiles) of the users absolute volume consumption per vendor. Notice the y-axis in logscale. For the sake of visibility, we only report on vendors with at least 1% of users and we group the remaining vendors as "other". The figure shows that Apple devices consume 3.6× and 1.6× (median values) more traffic than Microsoft and Android devices, respectively. If we focus on the fraction of HTTPS traffic (Fig. 4, right), we notice that the share of HTTPS is inversely proportional to the absolute volume, e.g., 50% of the Microsoft devices only consume about 2MB, out of which 90% is HTTPS.

We further investigate the HTTPS traffic generated by Microsoft devices and find that, on average, 60% of their traffic is addressed to Windows services like **.bing.** and **.live.**. A similar result holds for Motorola devices as well (having Google instead of Microsoft services). This suggests that most of this HTTPS traffic consists of "background noise", *i.e.,* communications generated by the operating system and apps but not strictly triggered by users activity. However, corroborating this belief with numbers is hard based on the available data.

Second Level Domain: Finally, we process the transaction hostnames to understand if they offer visibility on the HTTPS services. Recall that for HTTPS the hostname corresponds to the SNI communicated in the TLS handshake.

From each hostname, we remove the Top Level Domain (TLD) using the Mozilla Public Suffix list [7]. Then, for each Second Level Domain (SLD) found we compute the total number of bytes, and the associated share of HTTP and HTTPS. Overall, we find 1.6M SLDs, out of which 92% and 15% are used in HTTP and HTTPS transactions respectively. The heatmaps in Fig. 6 show the top-50 (left) and top-1000 (right) SLDs which account for 79.5% and 93.8% of volume in the whole day. SLDs mostly coincide with CDN providers; however, in some cases they accurately identify actual services (e.g., *streaming* – googlev-

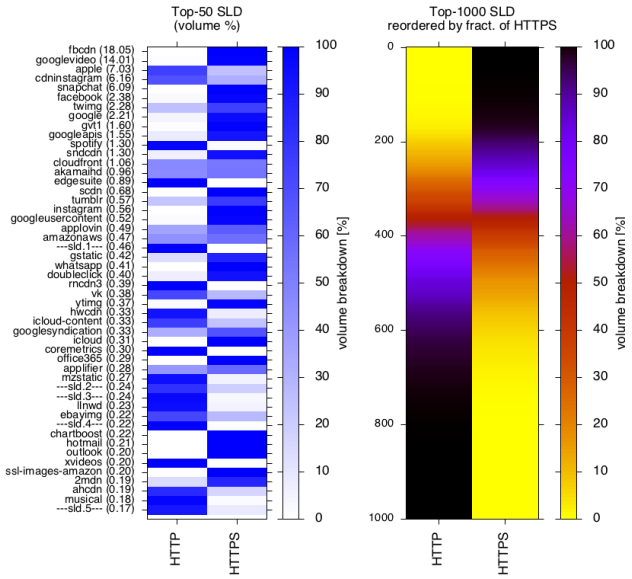

Fig. 6. Top Second Level Domain (SLD) volume breakdown.

ideo, spotify; *social network* – facebook, instagram, twitter, snapchat, whatsapp; *gaming* – applifier, etc.).

Takeaways*: We observed and quantified a "traffic gap" between HTTP and HTTPS. This gap originates from the different mix of services behind each protocol as well as OS (and device vendors). It follows that focusing on either one of the two protocols only for traffic-based analysis introduces a substantial bias not only on the overall volume, but also to capture device type diversity, and accessed services. A proper characterization of these aspects is key for mobile operators.*

5 Spatial and Temporal Analysis

Mobile operators are extremely interested in understanding *where/how/when* their customers consume data when moving across the network. This is crucial to drive investments (e.g., where to deploy/upgrade towers) and to support novel services such as geofencing and SON (Self Organizing Networks). In the literature, Call Data Records (CDRs) have been largely exploited to study human mobility [14]. However, a recent study [9] shows that UDRs (see Sec. 1) offer a richer vision on user mobility. We argue that web-logs enable an even finer grained spatial-temporal analysis than UDRs. This is because UDRs aggregate activities in (large) time windows, and associate them with coarse spatial information. Unfortunately, our dataset does not include UDRs and we thus cannot further quantify this intuition.

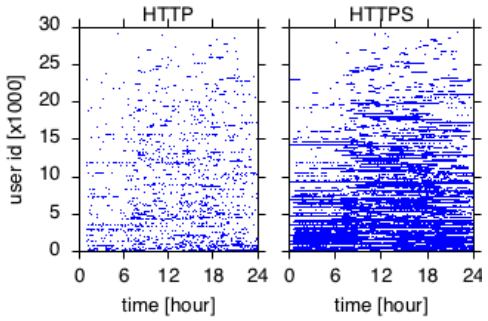

Fig. 7. HTTP and HTTPS activity over time (30k random users).

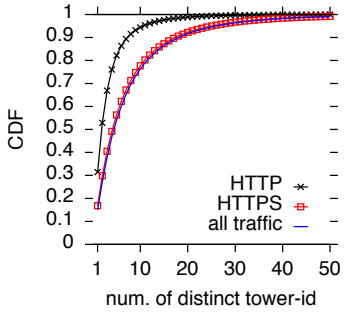

Fig. 8. Number of towers used (all users).

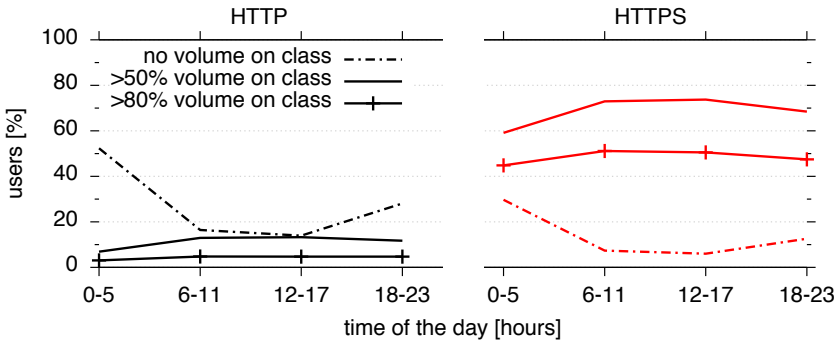

Fig. 9. Percentage of users consuming HTTP, HTTPS, and TCP-oth traffic with respect to time of the day.

We here explore to which extent HTTP traffic is representative of how users consume content across time and space when compared with HTTPS. As done in the literature [9], we approximate a user location with the position of the cell *towers* she is connected to. Note also that HTTP(S) transactions can be associated to different towers during their lifetime due to user mobility and/or load balancing at the radio layer. Our enrichment process (see Sec. 3) allows to identify all the towers associated to a transaction.

Traffic discontinuity: For each user we group HTTP and HTTPS transactions in 10 min bins. Each bin has a binary value depending if at least 1 transaction has been found or not. Fig. 7 shows the obtained bitmaps for 30,000 random users (results hold for other users). Notice how HTTP traffic (left plot) is more "discontinued" than HTTPS (right plot), *i.e.,* HTTP activity is more occasional and sparse across the day.

Fig. 9 shows a more detailed quantification of traffic variation across the day. We partition the day in 6 hours bins starting from midnight. Within each bin, we compute the percentage of users having 0%, >50%, or >80% of volume

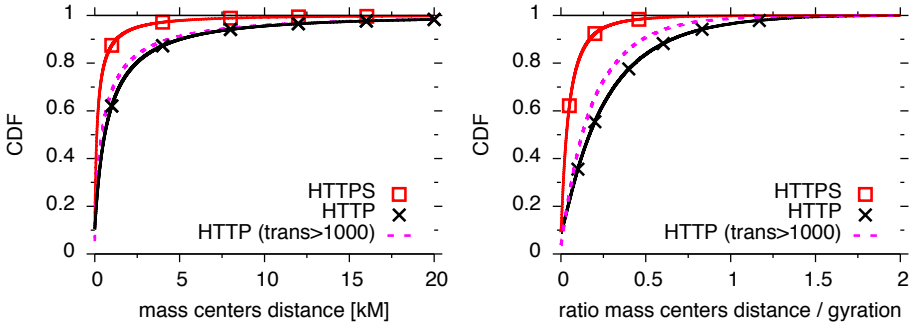

Fig. 10. Euclidean distance between mass centers (left) and normalized distance with respect to overall gyration (right).

over HTTP(S). Notice how HTTP better captures users activity during daily hours; instead, at night time (00:00–05:00) 50% of the users do not generate any HTTP traffic. As also observed for the device analysis (see Sec. 4), this suggests that HTTP traffic better captures real user activity rather than (automatic) background services.

Cell tower perspective: We here investigate how HTTP and HTTPS are consumed from a cell tower perspective. Fig. 8 reports the CDF of the number of distinct towers each device connect to during the day. When focusing on HTTP traffic only, we underestimate the set of towers contacted by a device. Specifically, only 6% of devices contact more than 10 towers, while such value doubles when focusing on HTTPS traffic only. The figure also shows that the HTTPS curve matches quite well the "all traffic" curve, which suggests that HTTPS is a very good "proxy" of the overall activity.

Next, we quantify how traffic is distributed among towers with the goal to identify per user "hot spots", *i.e.,* which towers carry most of the traffic for each user. We do this in term of number of transactions rather than volume as the presence of undetectable persistent connections in HTTPS can introduce a non-negligible error. Specifically, it is not easy to accurately split the volume of an HTTPS transaction across the towers it uses (see Sec. 3). We find that, for HTTP, 93% of the users consume at least 80% of their HTTP traffic in just 5 hot spots; this percentage reduces to 80% of the users when considering HTTPS. In term of hot spot similarity, we find a strong intersection: for 70% of the users, 7 out of 10 HTTP hot spots are also HTTPS hot spots. In other words, HTTP traffic alone seems to capture well the important locations where content is consumed.

Mass centers: To further corroborate on the previous result, we conclude our spatial analysis investigating "how distant in space" is HTTP traffic from HTTPS, and vice-versa. For each user, we compute a mass center [9] representing where HTTP, HTTPS, and the whole traffic is consumed. A mass center is computed as the average of towers coordinates weighted by their number of transactions. Let us call those points *mass-HTTP*, *mass-HTTPS*, and *mass-*

ALL respectively. We then compute the Euclidean distance between (*mass-ALL*, *mass-HTTP*) and (*mass-ALL*, *mass-HTTPS*) for each user. Fig. 10 (left) shows the CDF of the obtained distances. Results show that HTTP content tends to be consumed further away than the majority of the traffic. This result is independent from users activity, e.g., the plot obtained for very active HTTP users (trans> 1,000) is not significantly different from the plot obtained for the whole set of users.

To further quantify the "spatial gap" between HTTP and HTTPS, we normalize the euclidean distances with respect to the user *radius of gyration* computed considering the whole traffic activity. The gyration radius is a well established metric to characterize user mobility [9, 14]; it captures the average distance between a point (the mass center) and another set of points (all towers locations). For each user, we normalize the Euclidean distance between (*mass-ALL*, *mass-HTTP*) and (*mass-ALL*, *mass-HTTPS*) with the gyration obtained considering the whole traffic. Fig. 10 reports the CDFs of the normalized Euclidean distances. The figure shows that HTTPS captures very well user mobility, e.g., 97% of users have a normalized Euclidean distance < 0.5. The mobility observed from HTTP is very close to the overall one, and the normalized Euclidean distance tends to decrease for heavy HTTP users.

We conjecture that the latter result derives from the human component of the mobility problem. Users do not explicitly choose to use HTTP or HTTPS; the presence of a traffic type is an "artifact" of the device and applications used. However, a user chooses the location to visit and, as far as some traffic is consumed, this is enough to characterize her mobility pattern.

Takeaways: *The "time gap" between HTTP and HTTPS is substantial, with each protocol being respectively the most popular one at different points in time. Conversely, the "spatial gap" between HTTP and HTTPS is limited and both protocol are quite good in approximating user mobility.*

6 Conclusions

In this work we presented the first comparative study between HTTP and HTTPS traffic for mobile networks. The input of our study was a unique dataset including HTTP and HTTPS traffic, radio-layer information, and device information from a 10M-subscriber European mobile operator. Our analysis highlighted three different "gaps" between HTTP and HTTPS. First, a "traffic gap" related to how different services and OS/vendors use HTTP and HTTPS. Second, a "time gap" due to protocols being more used at different time of the day. Third a surprisingly small "spatial gap", probably motivated by the human component of mobility.

From an operator perspective, logging non-HTTP traffic implies additional investments. Based on the available dataset, we estimate that for a mid/large mobile operator such logging requires few additional TB of storage each day. This is an affordable investment compared to the potential need to upgrade the processing cluster along with the monitoring solution. Similarly, the extensive

adoption of persistent connections represent a hassle for monitoring network metrics that accurately reflect service performance. Overall, we argue that state of the art monitoring solutions are not yet ready to properly characterize HTTPS traffic. We hope that the results provided in this work quantify the importance of monitoring HTTPS, and they further stimulate the discussion in the research community towards creating better monitoring systems.

References

1. CAIDA: As rank, http://as-rank.caida.org
2. Casas, P., Fiadino, P., Bär, A.: Understanding HTTP traffic and CDN behavior from the eyes of a mobile ISP. In: Proc. Passive and Active Measurement Conference (PAM) (Mar 2014)
3. Erman, J., Ramakrishnan, K.: Understanding the super-sized traffic of the super bowl. In: Proc. ACM Internet Measurement Conference (IMC) (Oct 2013)
4. Erman, J.E., Gerber, A., Hajiaghayi, M., Pei, D.: To cache or not to cache: The 3g case. IEEE Internet Computing 15(2), 27–34 (2011)
5. Falaki, H., Lymberopoulos, D., Mahajan, R., Kandula, S., Estrin, D.: A first look at traffic on smartphones. In: Proc. ACM Internet Measurement Conference (IMC) (Nov 2010)
6. Keralapura, R., Nucci, A., Zhang, Z.L., Gao, L.: Profiling users in a 3g network using hourglass co-clustering. In: Proc. ACM MobiCom (Sep 2010)
7. Mozilla: Public Suffix List, http://publicsuffix.org/
8. Mucelli, E., Oliveira, R., Carneiro, A.V., Naveen, K.P., Sarraute, C.: Measurement-driven mobile data traffic modeling in a large metropolitan area. In: Proc. IEEE Conference on Pervasive Computing and Communications (PerCom). St. Luis, MO, USA (Mar 2015)
9. Ranjan, G., Zang, H., Zhang, Z.L., Bolot, J.: Are call detail records biased for sampling human mobility? ACM SIGCOMM Mobile Computer Communication Review 16(3), 33–44 (Dec 2012)
10. Sandvine, Global Internet Phenomena: Spotlight: Encrypted Internet Traffic, https://www.sandvine.com/trends/encryption.html
11. Shafiq, M.Z., Ji, L., Liu, A.X., Pang, J., Venkataraman, S., Wang, J.: A first look at cellular network performance during crowded events. In: Proc. ACM SIGMETRICS (Jun 2013)
12. Trestian, I., Ranjan, S., Kuzmanovic, A., Nucci, A.: Measuring serendipity: Connecting people, locations and interests in a mobile 3g network. In: Proc. ACM Internet Measurement Conference (IMC) (Nov 2009)
13. Vallina-Rodriguez, N., Sundaresan, S., Kreibich, C., Weaver, N., Paxson, V.: Beyond the radio: Illuminating the higher layers of mobile networks. In: Proc. ACM MobiSys (Nov 2015)
14. Vincent D. Blondel, Adeline Decuyper, G.K.: A survey of results on mobile phone datasets analysis. CoRR arXiv arXiv:1502.03406 (2015)
15. VNI, C.: The Zettabyte Era: Trends and Analysis, http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/vni-hyperconnectivity-wp.html
16. Xu, Qiang and Erman, Jeffrey and Gerber, Alexandre and Mao, Zhuoqing and Pang, Jeffrey and Venkataraman, Shobha: Identifying diverse usage behaviors of smartphone apps. In: Proc. ACM Internet Measurement Conference (IMC) (Nov 2011)