# Monitoring TLS Adoption using Backbone and Edge Traffic

Chia-ling Chan
Waseda University

Romain Fontugne
IIJ Research Lab

Kenjiro Cho
IIJ Research Lab

Shigeki Goto
Waseda University

*Abstract*—**Security and privacy have become major concerns for Internet communications. In order to prevent eavesdropping and tampering a lot of Internet protocols rely the Transport Security Layer (TLS). In this paper we aim to quantify the adoption of TLS using passive traffic traces captured on a backbone and edge academic network in Japan. We monitor the evolution of five common protocols and their TLS-variants over ten years of traffic data. We found that the adoption of TLS for HTTP really started around 2012, while IMAP traffic is mostly encrypted for the last ten years. The deployment of HTTPS is mainly driven by large content providers and migrating the remaining HTTP traffic to HTTPS might require significant efforts as it concerns numerous smaller services.**

## I. Introduction

The Internet was originally conceived to connect a limited set of honest networks for scientific purposes. Consequently, the original design of the Internet protocol suite mostly overlooked security and privacy issues. For example the Simple Message Transfer Protocol (SMTP) was designed to send emails but had initially no mean to authenticate senders or encrypt data.

Nowadays the Internet is central to most communications hence privacy and security have become major concerns. To secure end-to-end communications applications commonly rely on the Transport Security Layer (TLS), a cryptographic protocol that sits between the application protocol and the transport protocol. TLS prevents eavesdropping and tampering by encrypting the transmitted data and it also provides client/server authentication via digital certificates. Numerous protocols have now a *secure* variant that takes advantage of TLS. For example HTTPS is the adaptation of HTTP to TLS, it was originally designed for sensitive transactions, but it is now employed to protect any website from eavesdropping and ensure privacy and integrity of exchanged data. The benefits of HTTPS and other protocols using TLS are considerable but their deployment has been impeded by the added difficulty of obtaining and maintaining certificates.

In this paper, we aim to monitor the adoption of protocols using TLS using traffic traces collected at the backbone and edge of the Internet. We investigate the evolution of five different protocols for ten years of network traffic data.

We inspect closely the deployment of HTTPS has it represents the majority of the monitored traffic. We identify the Autonomous Systems that are responsible for the majority of HTTPS traffic and confirm that major content providers are the main driving force of HTTPS deployment.

For email traffic, we found that IMAP traffic is most entirely transmitted over TLS. However, SMTP is lagging behind other protocols in terms of TLS adoption.

We also inspected FTP traffic but found a very small number of TLS connections. As the SSH traffic is quite important in the analyzed traces we suspect users to prefer file transfer over SSH rather than using FTP's TLS variants.

## II. Background

In this study we analyze traffic traces captured at a backbone link and an access link for a university campus.

### A. Backbone Traffic: MAWI

The backbone traces come from the MAWI archive which is a traffic data repository maintained by the WIDE Project. The traffic is collected on a transit link between the WIDE network (AS2500) and an upstream provider. The archive contains daily packet traces (15 minutes in pcap format) from 2001 onwards. To ease computing time, in this paper we present results only for the $15^{th}$ of every month from January 2008 to August 2017, which represents about 681GB of pcap files.

### B. Edge Traffic: University Campus

The edge traffic is captured at the border router of a university campus in Japan from September 2014 until June 2017. This dataset is also composed of 15 minutes monthly traces and represents about 224GB of pcap files.

Although both dataset are captured at Japanese academic networks, the topological location and number of users is quite different. As shown in Section III and IV these differences have a certain impact on our results.

### C. Traffic Classification

To account for the fraction of traffic standing for a protocol or its TLS-variant we rely on a traffic classification tool called Libprotoident [1]. Libprotoident uses only the IP header, transport-layer header and the first four bytes of payload to find the application corresponding to a flow (i.e. packets with the same IP addresses, transport protocol and port numbers). It supports over 400 different applications, but for our study we retrieve only flows corresponding to the following applications: HTTP, SMTP, IMAP, POP3, FTP, and their corresponding TLS-variants.

To support TLS these protocols can adopt two different strategies, implicit TLS or STARTTLS. Implicit TLS means

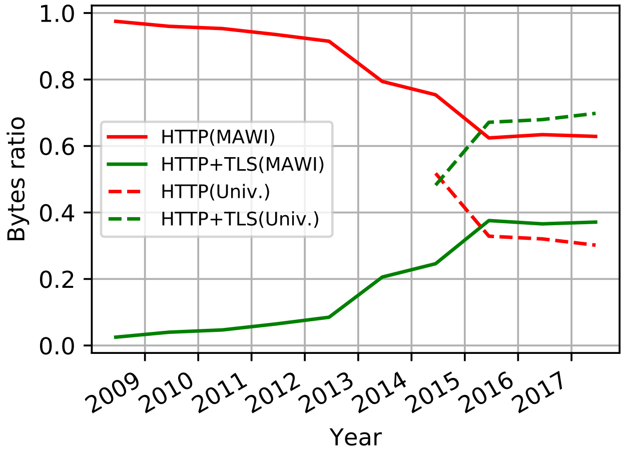

Fig. 1. **HTTPS adoption.** Relative number of bytes for HTTP and HTTPS as seen in MAWI (solid lines) and at the university campus (dashed lines).

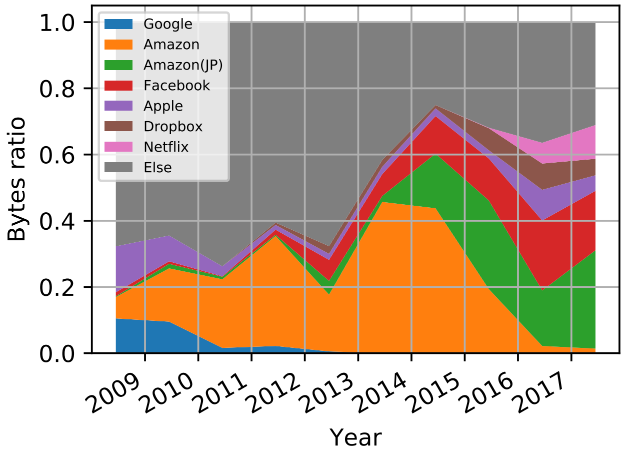

Fig. 2. **HTTPS: AS breakdown in MAWI.** Major autonomous systems observed for HTTPS traffic in MAWI.

that the connection starts with the TLS handshake then all commands from the application protocol are sent over the TLS connection. This requires that the connection starts on a port that is different from the unsecured-protocol, for example, HTTP connections are made on port 80 but HTTPS connections are via port 443. All the studied protocols have an implicit TLS variant deployed on the Internet, however, for email protocols and especially SMTP the STARTTLS variant is prevailing. A protocol implementing STARTTLS establish the connection using the usual port number and an application specific command will trigger the TLS handshake. Thereby STARTTLS does not require an additional port assignment but some modifications in the application protocol. Since all commands before the TLS handshake are sent in clear text the implicit TLS variant is usually recommended [2]. Because the STARTTLS command generally appears after a protocol handshake, which spans over the first four bytes monitored by Libprotoident, these STARTTLS-variants are not properly classified by Libprotoident. To circumvent this issue, we have modified Libprotoident to inspect payload in subsequent packets and detect corresponding STARTTLS commands. Our modified version of Libprotoident reports, for example, three types of application for the Internet Message Access Protocol: IMAP, IMAP_STARTTLS, and IMAPS. In the following we refer to both TLS-variants of a protocol using the notation, *protocol*+TLS. For example, IMAP+TLS, corresponds to both IMAP_STARTTLS and IMAPS.

## III. Deep dive into HTTP/HTTPS

Our analysis starts by looking at the adoption of TLS for the World Wide Web. This represents the vast majority of the monitored traffic, around 75% of the MAWI traffic (in terms of bytes) in 2017 is classified as web traffic. For all flows sent over port 80 or port 443, we compute the proportion of bytes that is transmitted over HTTP and the proportion for its TLS-variant, HTTPS, as reported by Libprotoident.

Figure 1 depicts the fraction of web traffic carried by the original protocol, HTTP, and the fraction sent over HTTPS for the two analyzed datasets. For MAWI in 2008, the original HTTP protocol accounts for 96% of the bytes transmitted with HTTP and HTTPS. From 2012 onward the adoption of HTTPS has significantly increased, reaching over 23% in 2014 and 36% in 2017. Some of the most recent traces in MAWI feature up to 50% of HTTPS traffic.

The HTTPS adoption for the university traffic is even more encouraging. We observe over 50% of the web traffic sent over HTTPS in 2014 and around 65% in 2017. For both datasets the HTTPS adoption is slowly increasing since 2015.

For web traffic (and all other applications, see Section IV) we found that the TLS adoption is much more prevailing in the university dataset. To understand this discrepancy between the two datasets we inspect the IP addresses corresponding to flows classified as HTTPS.

### A. ASN breakdown

We retrieve historical BGP data from the Route Views project and map monitored IP addresses to the corresponding Autonomous System Number (ASN) using longest prefix match. We discard ASNs from WIDE and networks that are downstream of WIDE, so we focus only on services located outside the WIDE network and its customer cone. This allows us to track the ASNs that are deploying TLS-enabled services and understand the adoption of HTTPS at a topological-level.

Figure 2 illustrates the relative number of HTTPS bytes per year for the most prominent ASNs in MAWI. Amazon is the main contributor for HTTPS traffic in this dataset. We also observed that the Amazon HTTPS traffic have migrated from AS14618 (referred as Amazon in Figure 2) to AS16509 which is another AS managed by Amazon with IP addresses located in Japan (referred as Amazon(JP) in Figure 2). We also found that Google was one of the main contributor for HTTPS before 2010, but then Google started peering directly with WIDE and

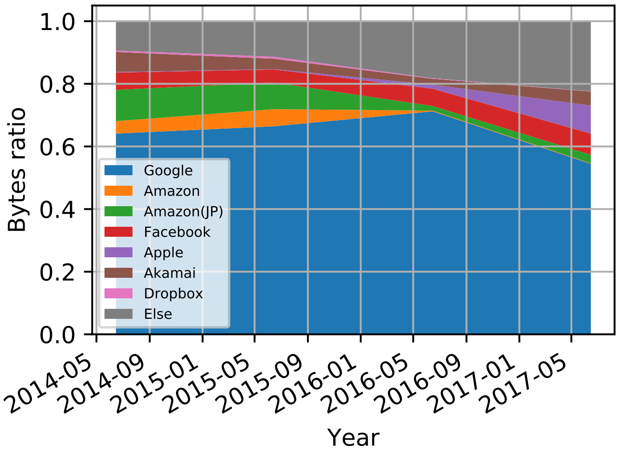

Fig. 3. **HTTPS: AS breakdown in university dataset.** Major autonomous systems observed for HTTPS traffic in the university campus.

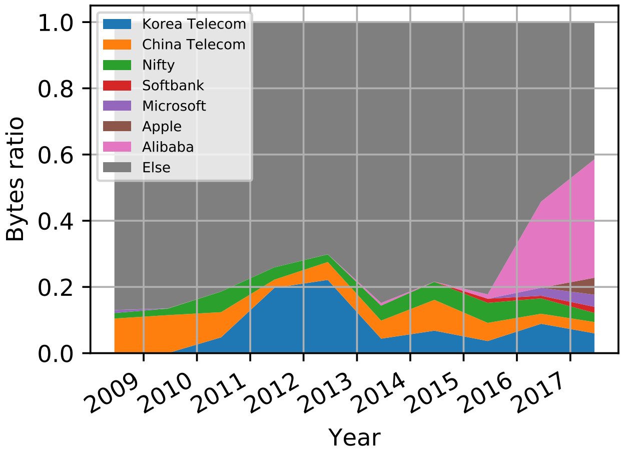

Fig. 4. **HTTP: AS breakdown in MAWI.** Major autonomous systems observed for HTTP traffic in MAWI.

less traffic was observed on the monitored transit link.

The top five ASNs (Amazon, Facebook, Apple, Netflix and Dropbox) accounts for over 65% of the total HTTPS traffic in 2017. Their share of HTTPS traffic have rapidly increased since 2010 when they represented less than 30% of the HTTPS traffic. By 2013, these five ASNs account for more than 50% of the total HTTPS traffic. The HTTPS traffic for Amazon is significantly growing from 2011 onward, HTTPS traffic for Facebook is growing steadily from 2012. Netflix arrived later in 2015 but its HTTPS traffic have grown very quickly.

When compared to Figure 1, the overall increase of HTTPS traffic occurs simultaneously with the HTTPS adoption of the major networks shown in Figure 2. Consequently, we argue that the large adoption of HTTPS is mainly thanks to the migration of large content providers to HTTPS.

The main difference with the university traffic is the presence of traffic to Google and Akamai (Figure 3). In this dataset Google alone represents over 60% of the monitored HTTPS traffic, which explains the discrepancy observed earlier between the MAWI and the university dataset (Figure 1). This observation emphasizes even more the role of large content provider in HTTPS traffic. Excluding Google traffic, the ASN breakdown is quite similar to the one obtained with the MAWI dataset. In 2017, Amazon, Facebook, Apple, represents about half of the HTTPS traffic excluding Google traffic.

For completeness, we also looked at the ASN breakdown for HTTP traffic. Figure 4 depicts the ASN breakdown for HTTP traffic observed in MAWI. The HTTP traffic to Alibaba (AS37963) has been impressively growing over the last couple of years. We also found that, unlike HTTPS, HTTP traffic is dispersed through numerous ASNs and most of the top ASNs are ISPs. Since HTTP traffic destinations are much more diverse than the ones for HTTPS, migrating the remaining HTTP web traffic to HTTPS (Fig.1) would likely require more effort than what have been accomplished so far. In addition, networks that have resources in countries where Internet censorship is very strict may be also discouraged to deploy HTTPS. For example, Wikipedia has been blocked in China when it turned HTTPS for all users in May 2015[1].

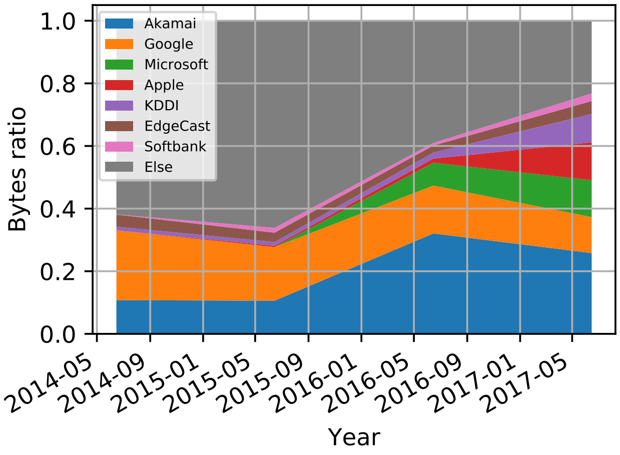

Fig. 5. **HTTP: AS breakdown in university dataset.** Major autonomous systems observed for HTTP traffic in university campus.

For the university dataset, Akamai is the main contributor of HTTP traffic, followed by Google and Microsoft (Figure 5). Despite Akamai stands out as one ASN, it represents numerous domains from Akamai's customers hence migrating all these domains to HTTPS may also require considerable efforts.

### *B. Country breakdown*

As previous studies reported a lower HTTPS adoption in Japan [3], we also investigated the geographical location of services found in HTTP traffic. Over the ten years of studied MAWI traffic about half of the observed HTTP traffic is

[1] https://www.theverge.com/2015/9/4/9260981/jimmy-wales-wikipedia-china

staying within Japan, around 25% is from U.S.A. and 10% is from China.

### C. Other protocols for web traffic?

The above results rely solely on the ability of Libprotoident to identify HTTP and HTTPS traffic. To validate our results we checked all results reported by Libprotoident for traffic on port 80 and 443 assuming that this ports are mainly, respectively, HTTP and HTTPS traffic. If Libprotoident reports a lot of unknown traffic for one of these ports it would indicate that Libprotoident misses some HTTP or HTTPS traffic, thus the above results might be biased.

Figure 6 depicts the applications found by Libprotoident for port 443 in the MAWI dataset. As expected we observe mainly HTTPS traffic on port 443. On average less than 5% of the bytes on port 443 are not classified as HTTPS. Over the entire studied period of time we observe sporadic HTTP traffic on port 443 and the QUIC protocol is emerging in 2015.

For the university traffic (Figure 7), the HTTPS traffic is also dominant, but we observe more QUIC traffic. This discrepancy between the two datasets is again explained by the lack of Google traffic in MAWI. In December 2015 and January 2016 we observed no QUIC traffic as Google disabled QUIC on their servers due to a vulnerability found in their implementation [4]. In 2017, QUIC represents about 10% of the bytes monitored on port 443 (TCP and UDP).

In our analysis we have disregarded QUIC as a TLS-variant of HTTP because this protocol is still under standardization and mainly used by Google. Considering QUIC as a TLS variant of HTTP slightly increases the TLS adoption mentioned above.

For port 443 Libprotoident reported very rarely unknown traffic and when it did it was mainly for UDP traffic (see Figure 6 and 7). We hypothesize that this traffic is also related to experimental HTTP protocols over UDP. The rare and infrequent appearance of this unknown traffic has no influence on our estimate of TLS adoption for HTTP.

We conducted the same verification for traffic on port 80. For both datasets we observed mainly HTTP traffic. Libprotoident reported significant unknown traffic ($>$ 20%) for one trace in 2015 and one in 2016. Overall we found that unknown traffic on port 80 is usually below 3%, which validates the results reported in previous sections.

## IV. Adoption of TLS for other applications

In this section our focus shifts to two common applications found in the analyzed traffic traces: email and file transfer. Standard protocols for these applications have all a TLS-variant that is also an IETF standard.

### A. Electronic Mail Traffic

Email is transferred via three different protocols. SMTP is designed for sending emails, it is essentially used to push emails to mail servers. IMAP and POP3 are conceived to retrieve emails and manage emails in a mailbox. POP3 offers less management functionalities hence it is slowly supplanted by IMAP.

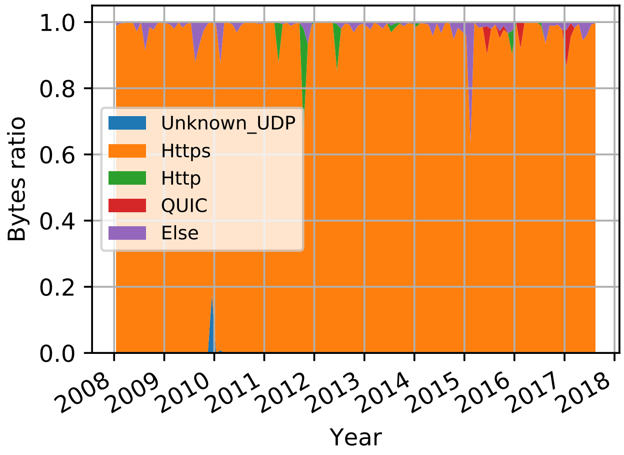

Fig. 6. Application breakdown for traffic on port 443 in MAWI.

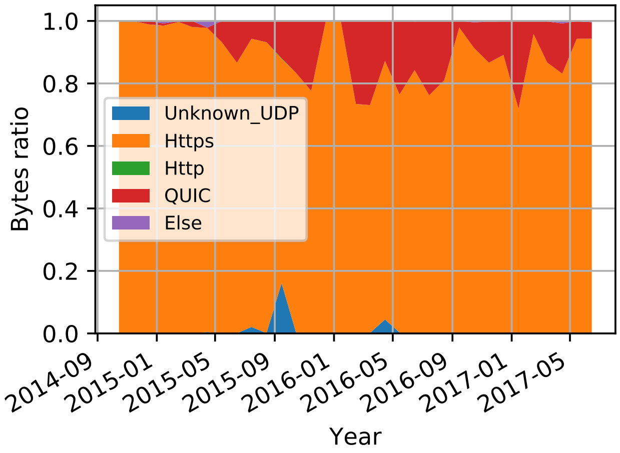

Fig. 7. Application breakdown for traffic on port 443 in the university dataset.

*1) IMAP:* Figure 8 depicts the fraction of traffic carried by the original protocol IMAP and its TLS-variants (IMAPS and IMAP with STARTTLS both referred as IMAP+TLS in Figure 8). For both datasets, IMAP+TLS accounts for almost all of the traffic. In 2016 for MAWI, IMAPS stands for 95%, IMAP with STARTTLS for over 4% and IMAP for less than 1%. We observe similar quantities for the university dataset. The change in 2014 is mainly due to an anomaly that happened on a single day.

*2) POP3:* Figure 9 depicts the fraction of traffic carried by the original protocol POP3 and its TLS variants (POP3S and POP3 with STARTTLS both referred as POP3+TLS in Figure 9).

For MAWI, as the POP3 traffic is not very important the ratios are quite unstable, but overall we observe a decreasing trend for the TLS-variants. This is surprising but the absolute values show that the amount of POP3 traffic is decreasing over

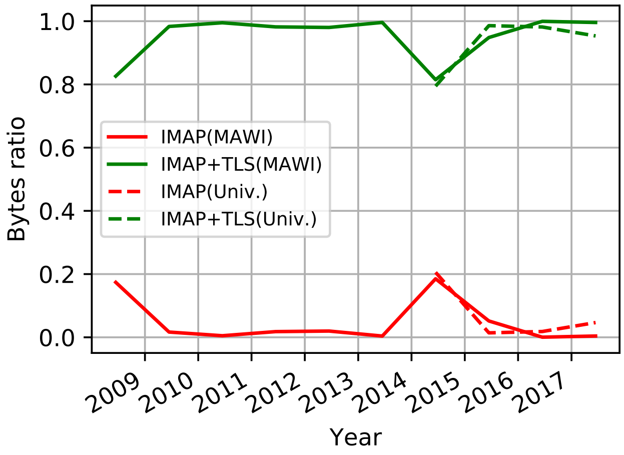

Fig. 8. **TLS adoption for IMAP.** Relative number of bytes for IMAP and IMAP+TLS (i.e. implicit and explicit TLS variants) as seen in MAWI (solid lines) and at the university campus (dashed lines).

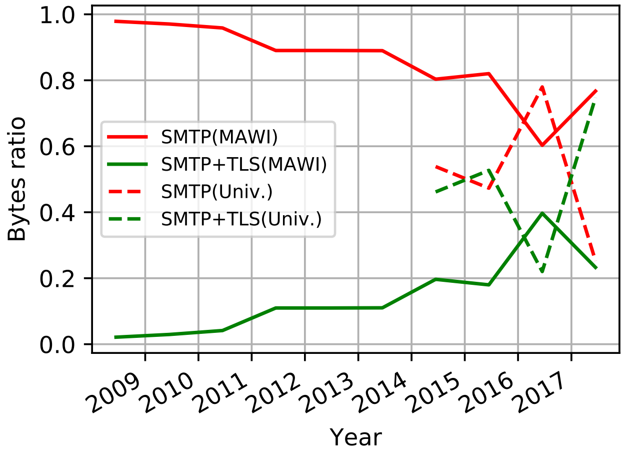

Fig. 10. **TLS adoption for SMTP.** Relative number of bytes for SMTP and SMTP+TLS (i.e. implicit and explicit TLS variants) as seen in MAWI (solid lines) and at the university campus (dashed lines).

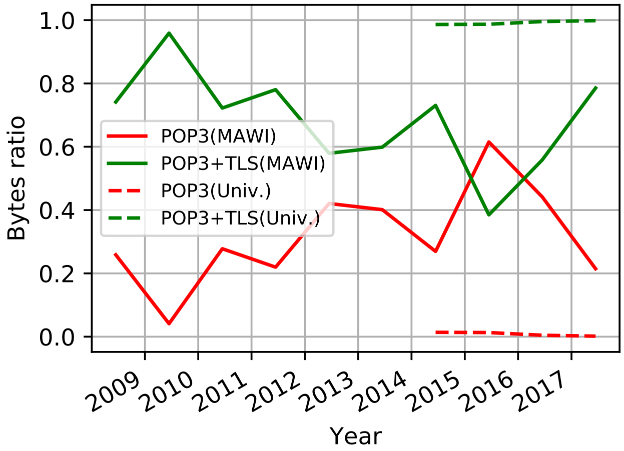

Fig. 9. **TLS adoption for POP3.** Relative number of bytes for POP3 and POP3+TLS (i.e. implicit and explicit TLS variants) as seen in MAWI (solid lines) and at the university campus (dashed lines).

time. Consequently, we conclude that users are moving away from POP3 and adopt either IMAP or web-based interfaces.

For the university dataset, most of the users are using POP3S. In 2017, 99% of the POP3 traffic in terms of bytes is transmitted through POP3S. Here again the two datasets highlight different results but we could not identify the cause of this discrepancy.

*3) SMTP:* Figure 10 depicts the fraction of traffic carried by the original protocol SMTP and it TLS variants.

In MAWI the adoption of TLS has been slowly increasing from 2009. In 2017 the ratio of TLS-variants is a little over 20% hence SMTP is one of the studied protocols that has the lowest TLS adoption.

For the university traffic, the ratio of the TLS variants are also usually lower than 50% although some very recent traces show more encouraging results.

For both datasets the STARTTLS alternative is much more popular that the implicit TLS variant on port 465. For example, in 2015 we observe about 20 times more traffic with the STARTTLS option than for the implicit TLS variant in MAWI (6 times more in the university dataset). This is expected as the implicit TLS variant on port 465 has been deprecated and the port is reassigned for another application. That said, traffic for the implicit TLS variant is not uncommon, we see such traffic through the entire study period for both datasets and some important surges in the university dataset in 2017. Recent discussions at the IETF also advocate for the use of the implicit TLS variant [2].

### *B. File Transfer*

We have also investigated the adoption of TLS for the FTP protocol. Although both the STARTTLS and implicit TLS variants exists for FTP, we very rarely see encrypted FTP traffic. We suspect that users prefer file transfers over SSH, as we observe more traffic over SSH that over FTP.

## V. Related Work

The adoption of TLS in the Internet has recently received a lot of attention. A recent study looked at the adoption of HTTPS using data gathered from two popular browsers, Google Chrome and Mozilla Firefox [3]. This study also shows that popular websites are more likely using HTTPS than other websites. It also reports a geographical disparities in HTTPS adoption, notably East Asia is lagging behind the rest of the world.

Durumeric et al. [5] employed several large-scale scans to study common characteristics of deployed HTTPS certificate. They also report a constant growth of domains supporting HTTPS from 2012 to 2013. These results are latter compared to other techniques in [6].

A few other studies investigated the use of TLS for other applications than web. Holz et al. [7] conducted Internet-wide scans and collected nine days of traffic to analyze the adoption of TLS for email and messaging protocols. Although major services have commonly adopted TLS, they report a globally low TLS deployment.

Finally, Carela et al. [8] studied the accuracy of six popular traffic classifiers and found that Libprotoident is the most accurate open source tool.

## VI. Discussion

### A. TLS trends

In contrast to previous works our study relies solely on passive monitoring techniques. Although active measurement techniques (e.g. Internet-wide scans) may provide a finer view of the global TLS deployment, passive monitoring enables us to focus on services that are favored by Internet users.

Overall we found that the adoption of TLS is increasing over time for all studied applications. The only exception is POP3 but the usage of this protocol is noticeably decreasing over the measurement period. In MAWI we observe from 2008 to 2017 a remarkable 14 times increase of the percentage of HTTPS traffic out of all web traffic (Figure 1).

The adoption of TLS is however varying significantly from one protocol to another. IMAPS is the most popular TLS-variant as it represents almost entirely the IMAP traffic observed in both datasets. On the other hand the adoption of TLS for SMTP is lagging behind other protocols. The FTP TLS-variants are almost never used but we found substantial traffic for SSH.

### B. Moving towards more HTTPS

HTTPS deployment is being successful thanks to the adoption from popular content providers, such as Google, Facebook and Amazon. Migrating the remaining HTTP traffic to HTTPS might however require more efforts as it concerns numerous services. Initiatives to ease the process of issuing certificates (e.g. Let's Encrypt[2] or Amazon AWS Certificate Manager[3]) are probably the best ways to enable anyone to adopt HTTPS.

We also suspect that the strict regulations of some countries may discourage local services to adopt HTTPS. For example, in 2017 we found that the majority of HTTP traffic corresponds to Chinese services. This observation also supports the lower HTTPS adoption reported in East Asia [3].

### C. Methodological takeaways

Our study relies on two sets of traffic traces captured on Japanese academic networks. Surprisingly we sometimes observed significant differences between the two datasets. Because of the recent lack of Google traffic in MAWI, HTTPS traffic in the university traces is relatively higher than the one in MAWI. Nonetheless, the HTTPS adoption in MAWI is closer to the one previously reported for Japan using browser data [3], hence the small population sample from the university dataset seems to be biased towards Google services. Consequently, we stress that such measurement study should employ multiple vantage points to understand the bias of each dataset and assess the validity of the results.

The metric used to quantify the adoption of a protocol is another difficulty we faced for this work. Like most past works, we employed the number of bytes to compare protocols usage. But when comparing a traffic intensive video service (e.g. Youtube or Netflix) to a simple text based service that have as many users, then the number of bytes is biased towards the former. Designing a more tailored metric to quantify the adoption of TLS would be very beneficial for assessing the advances of this protocol but this is a task we leave for future work.

## VII. Conclusion

In this paper we analyze passive traffic traces from a backbone and an edge network in Japan to measure the adoption of TLS for common Internet protocols. We found that the percentage of HTTPS over all web traffic has increased by a factor of 14 in the last ten years. We demonstrated that this trend is mainly due to the adoption of HTTPS by popular content providers. The remaining unencrypted HTTP traffic is mainly to services in Japan and East Asia. For email traffic we found that IMAP traffic is almost entirely transmitted over TLS, however, most of SMTP traffic is not using TLS. We then discussed the implications of our results and some methodological challenges.

## Acknowledgments

We thank Randy Bush, Simone Ferlin, Kazuhiko Yamamoto, and Oliver Gasser for the discussions and their valuable comments during the early phases of this work.

## References

[1] Shane Alcock and Richard Nelson, “Libprotoident: traffic classification using lightweight packet inspection,” *WAND Network Research Group, Tech. Rep.*, 2012.

[2] Keith Moore and Chris Newman, “Cleartext Considered Obsolete: Use of TLS for Email Submission and Access,” Internet-Draft draft-ietf-uta-email-deep-12, Internet Engineering Task Force, Dec. 2017, Work in Progress.

[3] Adrienne Porter Felt, Richard Barnes, April King, Chris Palmer, Chris Bentzel, and Parisa Tabriz, “Measuring HTTPS adoption on the web,” .

[4] Adam Langley, Alistair Riddoch, Alyssa Wilk, Antonio Vicente, Charles Krasic, Dan Zhang, Fan Yang, Fedor Kouranov, Ian Swett, Janardhan Iyengar, et al., “The QUIC Transport Protocol: Design and Internet-Scale Deployment,” in *Proceedings of the Conference of the ACM Special Interest Group on Data Communication*. ACM, 2017, pp. 183–196.

[5] Zakir Durumeric, James Kasten, Michael Bailey, and J Alex Halderman, “Analysis of the HTTPS certificate ecosystem,” in *Proceedings of the 2013 conference on Internet measurement conference*. ACM, 2013, pp. 291–304.

[6] Benjamin VanderSloot, Johanna Amann, Matthew Bernhard, Zakir Durumeric, Michael Bailey, and J Alex Halderman, “Towards a complete view of the certificate ecosystem,” in *Proceedings of the 2016 ACM on internet measurement conference*. ACM, 2016, pp. 543–549.

[7] Ralph Holz, Johanna Amann, Olivier Mehani, Mohamed Ali Kâafar, and Matthias Wachs, “TLS in the Wild: An Internet-wide Analysis of TLS-based Protocols for Electronic Communication,” in *NDSS*. 2016, The Internet Society.

[8] Valentín Carela-Español, Tomasz Bujlow, and Pere Barlet-Ros, “Is Our Ground-Truth for Traffic Classification Reliable?,” in *PAM'14*. Springer-Verlag New York, Inc., 2014, pp. 98–108.

[2] https://letsencrypt.org

[3] https://aws.amazon.com/certificate-manager/