Measuring HTTPS Adoption on the Web again

Christoph Werner
Matr.: 01529111
Curriculum: 033 534
e01529111@student.tuwien.ac.at

February 14, 2019

Abstract

This paper gives an overview of HTTPS, a mechanism to ensure privacy and integrity on the Web, and then tries to reproduce the HTTPS support among different set of websites as done in the paper "Measuring HTTPS Adoption on the Web" by Adrienne Porter Felt, Richard Barnes, April King, Chris Palmer, Chris Bentzel and Parisa Tabriz. The results will then be evaluated and compared to the results of the paper.

1 HTTP and HTTPS

HTTP (Hypertext Transfer Protocol) and HTTPS (Hypertext Transfer Protocol Secure) are two common protocols that often occur when talking about the Web. Both of them will now be shortly described.

1.1 HTTP

HTTP is the basic protocol powering the web. When a client connects to a server requesting a resource the server responds with status information and the resource if available as described in the HTTP protocol. The client can then evaluate the response and based on the information can take further actions like sending more requests to the server requesting further resources. The HTTP protocol is stateless so information from previous requests are lost. Every request is independent from the previous ones.

The main disadvantages of HTTP are the missing security features. An attacker can easily read and edit the intercepted traffic. This means that the client and the server can never be sure that their communication isn't read or edited by third parties. A further disadvantage is that the client can not be sure that the server he communicates with is the one he wants to communicate with. An attacker could fake the real website and trick the user into submitting sensitive information to his fake site.

1.2 HTTPS

Because of the disadvantages of the HTTP protocol the secure version HTTPS has been developed. Instead of directly transferring the plain HTTP messages, the messages are first cryptographically secured with Transport Layer Security (TLS). This protects the HTTP messages from being read and edited giving the client and the server the confidentiality and integrity of

the data.

A further security mechanism of the HTTPS protocol is the usage of digital certificates. Certificates are given out by Certificate Authorities (CA) to website owners. This ensures the authenticity of the server the client is in contact with. This gives the user some sort of protection against attackers faking websites.

All this security mechanisms of the HTTPS protocol make the web a safer place for users but does primarily protect against network attacks. Other classes of attacks like attacks on the web-level are still possible and put the users at risk. An example are Cross Site Scripting (XSS) attacks that can be prevented by Content Security Policies (CSP). Mechanisms like Content Security Policies are however dependent on HTTPS to be secured against attacks on the network layer because otherwise protections like these are rendered useless. [3]

HTTPS is therefore a base requirement for security on the internet.

2 The tests

Looking at the 4th section of the "Measuring HTTPS Adoption on the Web" paper it can be seen that different tools have been used for different lists for measuring if HTTPS is available as well as if HTTPS is the default. Also the definitions for HTTPS available and HTTPS default differ depending on the tool used. For example HTTPS default is differently defined in the Google Transparency Report and in HTTPSWatch. In the Google Transparency Report it is not necessary that the HSTS header is set. In HTTPSWatch it is required that the header is set. [3] With the HSTS (HTTP Strict Transport Security) header the server can instruct the browser that further connections should only be made with HTTPS to avoid downgrade attacks.

2.1 HTTPSWatch

HTTPSWatch (`https://httpswatch.com`) is a site that lists several popular websites for several categories and tests them for their HTTPS support. There are three different levels of HTTPS quality defined on the website. These are "Bad", "Mediocre" and "Good". The "Bad" rating is given if it is not possible to establish a TLS connection or no page can be transmitted over TLS. "Mediocre" means that there are quality issues like a missing HSTS header or no automatic redirect to the HTTPS version of a website. The last rating "Good" is given if everything is ok.[2]

The "Measuring HTTPS Adoption on the Web" paper defines HTTPS available for a rating with "Mediocre" or "Good" and HTTPS default for a rating with "Good". [3] There is one site with a "Bad" rating, 17 sites with a "Mediocre" rating and 21 sites with a "Good" rating. Based on these numbers this means that there are 38 sites where HTTPS is available and 21 sites where HTTPS is the default of 39 sites in total.

It is notable that the list of websites on HTTPSWatch has changed. For example the website `vine.co` can no longer be found on HTTPSWatch. Also there might be a website missing in the archived list of HTTPSWatch or the count of websites in the "Measuring HTTPS Adoption on the Web" is wrong because the list size in the archived list content of HTTPSWatch

of the paper only states 39 websites while in the table a list size of 40 is mentioned for the HTTPSWatch result.

2.2 Google Transparency Report

In the Google Transparency Report (`https://transparencyreport.google.com/https/top-sites`) Google scans a set of 100 non-Google websites that according to Google accounts for approximately 25% of all website traffic worldwide.[1] Although in the "Measuring HTTPS Adoption on the Web" paper is stated that the results are updated weekly the last scan is from the 20th October 2017. Due to the fact that the Googlebot is proprietary software the scans can't be reproduced exactly the same. For further measurements and replications regarding the Google Transparency Report the Mozilla Observatory and the data in the archived list "A.3 Google Transparency Report" of the paper is used. It is notable that the definition for "HTTPS available" of the Mozilla Observatory is different. While in case of the Google Transparency Report "HTTPS available" requires the Googlebot to not being redirected over an HTTP location the Mozilla Observatory on the other hand ignores redirections.[3] Due to this differences the results can't be exactly compared but gives a good indication on how the state of HTTPS changed compared to the data of the Google Transparency Report.

Of a total from 100 scanned websites 87 websites have HTTPS available. To one website "haosou.com" the request failed overall and a total of 40 websites automatically redirect to a HTTPS website.

2.3 Alexa Top 100 Global

The Alexa Top 100 are the 100 most popular websites based on the Alexa traffic estimates. The data can be fetched from Amazons AWS but not free of charge. Therefore only the current Top 100 Global are considered in this paper due to too high costs for fetching the top million websites as it has been done in the "Measuring HTTPS Adoption on the Web" paper.

The dataset can be queried from Amazons AWS (see `https://aws.amazon.com/alexa-top-sites/` for further information).

A total of 87 websites of the Alexa Top 100 Global have HTTPS available and 35 also automatically redirect to HTTPS in case a HTTP site has been requested. For this dataset also one request to a site failed. The request to "microsoftonline.com" has not been successful even though websites with valid HTTPS certificates are delivered on subdomains. Such special cases are not considered further and scans like these are considered to just have failed.

2.4 Censys

Censys is a company that scans the internet and provides a public search engine to query different information about the hosts and networks that make up the internet. It also provides a large database of server configurations with information about the TLS support.[3]

The data can be directly queried on the Censys homepage `https://censys.io`. For fetching the number of websites with HTTPS the search query

```
443.https.tls.validation.browser_trusted: true and
```

```
protocols: "443/https" and
443.https.tls.validation.matches_domain: true
```

has been used. This results in a total of 933,580 websites at the time of writing that match the specified search parameters. Without any search parameters the search engine finds a total of 1,431,351 websites.

For the check of a valid certificate for IPV4 hosts the last check for a matching domain must be removed as already described in "Measuring HTTPS Adoption on the Web". [3] So the final query is

```
443.https.tls.validation.browser_trusted: true and
protocols: "443/https"
```

with the "IPV4 Hosts" dataset selected. A total of 109,147,010 hosts could be found with 21,543,477 having HTTPS available.

The queries have been performed on January 19, 2019.

3 Results and comparison

Asking the same question as in the "HTTPS Adoption on the Web" paper "Is HTTPS support increasing?" can clearly be answered with "Yes". As it can be seen in the tables 1 and 2 the HTTPS availability increased as well as the default HTTPS distribution. While many popular websites already deploy HTTPS and only a small subset do not, in regard to defaulting to HTTPS a lot more can be done. The number of default HTTPS increased but is still very low which puts users at risk.

List	**List size**	**Tool**	**HTTPS available**	**Default HTTPS**
HTTPSWatch Global	40	HTTPSWatch	80%	35%
Google Top 100	100	Googlebot	54%	44%
Alexa Top 100 Global	100	Mozilla Observatory	87%	23%
Alexa Million	969,278	Mozilla Observatory	40%	10%
Alexa Million	856,312	Censys	38%	N/A
IPv4 hosts	101,052,620	Censys	10%	N/A

Table 1: Results from "Measuring HTTPS Adoption on the Web"

List	**List size**	**Tool**	**HTTPS available**	**Default HTTPS**
HTTPSWatch Global	39	HTTPSWatch	97%	54%
Google Top 100	100	Mozilla Observatory	87%	40%
Alexa Top 100 Global	100	Mozilla Observatory	87%	35%
Alexa Million	1,431,351	Censys	65%	N/A
IPv4 hosts		Censys	19%	N/A

Table 2: New results of HTTPS adoption

4 Repeatability

The scans done in this paper can easily be repeated. All lists of websites that have been used in the analysis can be found at the end of the paper (see appendix A and B) except of the list

of the Alexa Top 100 Global which can be queried from Amazons AWS. It is notable that the Alexa Top 100 Global list is subject to change and so the returned domains might differ to those that were used in this paper for the scans when using the current top 100 sites.

For running the Mozilla Observatory tests in an automated way the code at `https://github.com/chrztoph/measuring-https-adoption-on-the-web-again` can be used. Further instructions on how to use the tool can be found there.

References

[1] Google Transparency Report Top Sites. `https://transparencyreport.google.com/https/top-sites`. Accessed: 2018-12-31.

[2] HTTPSWatch About. `https://httpswatch.com`. Accessed: 2018-12-31.

[3] April King Chris Palmer Chris Bentzel Parisa Tabriz Adrienne Porter Felt, Richard Barnes. Measuring HTTP Adoption on the Web. 2017.

Appendices

A HTTPSWatch List

www.baidu.com, www.bing.com, duckduckgo.com, www.google.com, www.sohu.com, www.yandex.ru, www.yahoo.com, www.linkedin.com, www.facebook.com, www.twitter.com, www.pinterest.com, instagram.com, www.reddit.com, www.youtube.com, vine.co, www.match.com, www.okcupid.com, disqus.com, store.apple.com, www.amazon.com, www.bestbuy.com, www.ebay.com, www.craigslist.org, www.target.com, www.walmart.com, www.cvs.com, www.homedepot.com, www.barnesandnoble.com, www.box.com, www.dropbox.com, drive.google.com, www.icloud.com, onedrive.live.com, www.tarsnap.com, www.blogger.com, medium.com, squarespace.com, staff.tumblr.com, wordpress.com

B Google Transparency Report

360.cn, aliexpress.com, amazon.co.jp, amazon.co.uk, amazon.com, amazon.de, amazon.in, apple.com, ask.com, ask.fm, beeg.com, bongacams.com, chaturbate.com, cnet.com, craigslist.org, ebay.co.uk, facebook.com, fc2.com, forbes.com, goo.ne.jp, haosou.com, imgur.com, instagram.com, linkedin.com, mail.ru, netflix.com, nih.gov, nytimes.com, ok.ru, onet.pl, paypal.com, pinterest.com, pornhub.com, pzy.be, rakuten.co.jp, reddit.com, redtube.com, seznam.cz, softonic.com, stackoverflow.com, taobao.com, theguardian.com, tmall.com, tripadvisor.com, tumblr.com, twitch.tv, twitter.com, vk.com, whatsapp.com, wikihow.com, wikimedia.org, wikipedia.org, wordpress.com, wp.pl, xhamster.com, xnxx.com, xvideos.com, yahoo.co.jp, yahoo.com, yandex.ru, yelp.com, youporn.com, baidu.com, cnn.com, ebay.com, sohu.com, t.co, alibaba.com, amazonaws.com, bbc.co.uk, bing.com, chinadaily.com.cn, dailymail.co.uk, dailymotion.com, daum.net, globo.com, gmw.cn, go.com, goal.com, hao123.com, imagebam.com, imdb.com, live.com, microsoft.com, milliyet.com.tr, mirror.co.uk, msn.com,

naver.com, naver.jp, office.com, olx.biz.id, qq.com, sina.com.cn, soso.com, telegraph.co.uk, tianya.cn, uol.com.br, weibo.com, wikia.com, xinhuanet.com