Survey of United States Related Domains: Secure Network Protocol Analysis

DeJean Dunbar, Patrick Hill, and Yu-Ju Lin

Department of Computer Science,
Charleston Southern University, North Charleston, South Carolina

ABSTRACT

Over time, the HTTP Protocol has undergone significant evolution. HTTP was the internet's foundation for data communication. When network security threats became prevalent, HTTPS became a widely accepted technology for assisting in a domain's defense. HTTPS supported two security protocols: secure socket layer (SSL) and transport layer security (TLS). Additionally, the HTTP Strict Transport Security (HSTS) protocol was included to strengthen the HTTPS protocol. Numerous cyber-attacks occurred in the United States, and many of these attacks could have been avoided simply by implementing domains with the most up-to-date HTTP security mechanisms. This study seeks to accomplish two objectives: 1. Determine the degree to which US-related domains are configured optimally for HTTP security protocol setup; 2. Create a generic scoring system for a domain's network security based on the following factors: SSL version, TLS version, and presence of HSTS to easily determine where a domain stands. We found through our analysis and scoring system incorporation that US-related domains showed a positive trend for secure network protocol setup, but there is still room for improvement. In order to safeguard unwanted cyber-attacks, current HTTPS domains need to be extensively investigated to identify if they possess lower version protocol support. Due to domains supporting lower HTTPS protocol support, there needs to be further support for browsers enforcing the most up to date protocol version. Additionally due to the infrequent occurrence of HSTS in the evaluated domains, the computer science community necessitates further HSTS education by introducing the basic concepts in classroom settings across the United States

KEYWORDS

Network Protocols, HTTP Strict Transport Security, scoring benchmark, domain analysis, survey

1. INTRODUCTION

HTTP was a pinnacle basis in defining how information was transmitted across a network during this technical era of information technology. Since 1990, the world wide web has employed the

Commented [STH1]: I could not find a problem statement or even a thesis statement in this introduction.

Commented [dd2R1]: Adding problem statement and thesis statement in the 2nd to last paragraph in Introduction section

HTTP Protocol as a stateless application-level protocol for hypermedia information systems. HTTP/0.9, the initial implementation of HTTP, was designed for raw data delivery across the Internet. As new versions of HTTP were released, no security procedures for the transport of raw data were implemented [1]. Hackers can access health information, government information, and personal information. HTTPS protocol was introduced to address this significant security vulnerability.

HTTPS is a direct extension of the HTTP Protocol introduced by Netscape Communications. There were several security implementation versions of the HTTPS protocol which were SSL and TLS. SSL was first proposed in the middle of 1994 by Netscape Communications with its highest version implementation being version 3. TLS was proposed by the Internet Engineering Task Force with its highest version being version 1.3. The overall goal of HTTPS was to provide security features for HTTP such as encipherment, digital signature mechanisms, data integrity, authentication exchange mechanisms, and notarization mechanisms [2]. With these additional security measures incorporated, another web security standard extension for HTTPS was introduced known as HTTP Strict Transport Security (HSTS).

The Internet Engineering task force (IEFT) proposed HSTS in 2012 and defined it as a security mechanism that restricts website access to only secure connections. This feature guards against bootstrap man in the middle (MITM) attacks. The MITM attack allows hackers to interfere in the communication between two systems and has been proven that it can occur even in domains with HTTPS protocol [18]. Additionally, HSTS provides security by converting a URI reference to a secure URI reference [3]. With these additional security benefits that can be added to HTTPS, the HSTS security mechanism helps a domain's network security strength to be even stronger.

Despite network advances in security, cyber-attacks are still prevalent in the United States which creates a high need to ensure proper security exists for websites. It is important to acquire a sense

of where United States managed domains stand currently by checking out its network protocol setup components such as HTTPS protocol version and whether it contains an HSTS header.

This research is structured with related works in section two which discusses past research efforts in domain analysing ,scoring benchmarks, and HSTS trends. Section three presents a methods segment consisting of the hardware involved, techniques for gathering the US related domains, the database used, the scanner used for domain protocol information , python parsers used, the proposed scoring system, and information about the HSTS survey. Section four addresses the experiment results/analysis which includes considerations made in the research gathering, research results, and limitations. Section five follows with the impacts of the research. Finally, section six ends with the conclusions based off the research results.

2. LITERATURE REVIEW

2.1. Past Domain Analysing

This segment showcases past research effort in the category of domain scanning. The work summarizes and provides an explanation of how the it contributes to the overall research in this paper.

Patrick Hill's and Yu-Ju Lin's research establishes the analyzing of government websites in the United States due to the trustworthiness of state and local government websites being questionable. They concluded that there are instances of government websites that are not as secure as they should be against cyber-attacks [12]. However, the methodology used to collect the ".gov" domains in the other work included an exhaustive search through Google for the top ten US counties, which may have left out several domains. To create a more accurate assessment, this research will use a defined list of all ".gov" domains by obtaining a complete list from the registrar DOTGOV website. Patrick's Hill's and Yu-Ju Lin's research impacts this current research because it prevented a very tedious and unreliable way of gathering ".gov" domains. Another impactful way their research connects with my current research is that the inspiration of

analyzing other United States based entities came from their thorough and detailed focus on ".gov" domains.

Another research paper focuses on the evaluation of many Arabic-language websites. The procedure for determining what was considered Arabic consisted of two steps. The author created 16 search terms by utilizing Google Trends for a specified year range. Following that, the search terms are submitted to Google and all URLs returned in the results are scanned [17]. Our research will rely on simple methods such as acquiring a zone file for ".us" domains, utilizing a trusted web crawling website for ".edu" domains, and utilizing the government website that hosts all ".gov" websites in the United States.

Another study focuses its research on scanning over 22,000 websites originating in Europe to determine their level of security [18]. The researcher filtered out European domains in this study by using country code top-level domains associated with Europe, such as ".be" (Belgium). In our research, we will replicate this approach for scanning websites related to the United States by utilizing the top-level domain ".us" as a source for domain analyses.

An additional work focuses on measuring HTTPS usage on the internet [21]. The author is primarily concerned with how far HTTPS has progressed in its adoption on the web. One important aspect of the paper was the assurance that the security community has invested in supporting and requiring HTTPS through related works such as the automated certificate authority known as "Let's Encrypt," the non-profit SSL Labs, and technologist moving government websites to HTTPS. This work contributes to the current research because it can potentially supplement the results of our work when it comes to the current progress of HTTPS adoption of the domain groups chosen in the United States.

2.2. Past Scoring Benchmark

It is necessary to review prior research that has been conducted on scoring systems. The goal of this section is to give a brief overview of the related entity along with providing how our research makes a contribution.

SSL Labs is a non-profit research organization that maintains a comprehensive collection of SSL/TLS documentation, tools, and a community [11]. Their website is capable of scanning domains for SSL configurations and providing a score. The overall SSL/TLS strength score for a domain is divided into three categories: protocol support, key exchange, and cipher strength. SSL Labs' current shortcoming is that they have not included TLS 1.3 in their rating guide. Our research will contribute by incorporating a scoring system that incorporates TLS 1.3 as well as determining whether the domain supports HSTS. This related work had an impact on my research because it provided a general concept of how to determine scoring for protocol versions, which led me to focus my proposed scoring benchmark on protocol support as a scoring factor. Another source of research for past scoring benchmark included studying website security response headers in an Indonesia Bank. The author's scoring system used in this research was based on characteristics that exists for a header element [19]. The scoring system the author composed was effective for response header characteristics, however our research takes a mixed approach checking network protocol version along with the response header characteristic of HSTS presence.

2.3. HSTS Header Past Trends

It is important to review past studies of HSTS presence in websites. The purpose of this section is to provide a brief overview of these trends so that they can be used as a comparison later in the research to see if the occurrence of HSTS has increased in a positive trend.

The study of occurrence of HSTS in domains is supported by other works conducted in the past. One research focused on government websites had a trend in where only 2.86% percent of those websites supported HSTS [12]. Another research paper on a HSTS deployment survey conducted in 2013 revealed a trend for HSTS. Of the 1 million websites analysed in this research only 277 contained HSTS headers [13]. Additional research work conducted in 2018 focused on analysing the adoption of security headers in HTTP found that from the 1 million websites scanned that only 5.41% used HSTS [14]. Another study conducted in July 2018 focused on analysing HSTS

found that from the 1 million websites scanned that only 5.35% used HSTS [15]. Another author's research focused on scanning a list of domains for the presence of HSTS. The author conducted web crawls using Alexa's top million website list. The author determined from their analysis that only 1.1% of the websites are setting HSTS headers [16].

Commented [STH3]: Move to Chapter 2.

Previous research in domain scanning , scoring benchmarks, and HSTS occurrence paved the way for the success of this study. These works established a foundation for domain gathering techniques scoring benchmarks, and trends of HSTS occurrence. However, some preliminary work is required before these concepts can be implemented and revealed in this research.

3. METHODS

This section of the study goes through the hardware used in the experiment. The origins of domains related with the United States is also highlighted. Next the database housing the domains is revealed. Following that, the scanning procedure used to detect the domain's SSL/TLS version is focused on . Additionally, the numerous Python scripts used to help with domain analysis and database updates are described . Finally, the scoring benchmark proposed in this study will be explained in detail.

3.1. Hardware Used

All the tools used for this research were all run in a virtual machine using Ubuntu. The computer model is an Inspiron 16 7610 which operates on a Windows 11 Pro x64 operating system. The computer has an i7 core and 32GB RAM. It was necessary to utilize a higher core to better utilize throughput for running multiple instances of the scanning program and the python scripts.

3.2. Techniques for Gathering US Based Domains

For this study, domains ending in ".us", ".gov", and ".edu" were grouped together. One reason is that the theme of United States is involved in all three domains. For example, ".gov" domains are reserved for US government entities; ".edu" domains are reserved for US educational institutions; and ".us" domains are reserved for US citizens and entities. This gives high confidence that these domain types being analyzed are United States based domains. The number of ".us" domains

Commented [STH4]: First mention of .us and .edu. Why are these included? Why just these? Justify this choice.

Commented [dd5R4]: Added reasoning to support the choices

gathered were 1,814,204. The number of “edu” domains gathered were 5,854. The number of “.gov” domains gathered were 7,671. The grand total of domains gathered were 1,827,729. In order to study how the domain groups were gathered, we first focus on the technique for gathering “.us” domains.

Commented [STH6]: These are conflicting numbers for .gov. What about .edu?

Commented [dd7R6]: Fixing this sentence as there seems to be mention of .gov twice when it should have been .us , .gov and .edu

To find the source of “.us” domains, the first step was to determine what entity had access to the zone file. A “.us” zone file request was sent to the registry site ABOUT and was later redirected to GODADDY [4]. A zone file contains a list of all domains that have been registered. Following the approval of the request, the zone file was provided. A new zone file is created every day with the year, month, and day due to new/existing domains being updated. The zone file we chose for parsing was from March 26,2022. Although domain names were included in the zone file, the file contained other information deemed unimportant. On a DELL laptop running Ubuntu virtually, a series of commands shown in Figure 1 was executed on an Ubuntu command prompt to extract only the domain names from the zone file. The extracted data was saved as text files.

```
$ awk '{print $1}' us.zone > domains-only.txt
$ sort- u domains-only.txt --output domains-unique.txt
$ LC_ALL=C grep '^[A-Z0-9\-]*$' domains-unique.txt > domains.txt
```

Figure 1. List of commands used to parse through “.us” zone file

The same technique was attempted for “.edu” domains, however the organization EDUCAUSE’s cooperative agreement rules would not allow them to give us access to their zone file. To accommodate the lack of a zone file, we decided to utilize two outside sources for the gathering of “.edu” domains. The first source was from a GitHub repository which provided “.edu” domains of universities from around the world [5]. The list was filtered to only the United States. The next source of ".edu" domains came from Common Crawl, a reputable web crawling service [6]. Common Crawl provided a server for their data to be queried via Amazon's AWS service, Athena [7]. A query was run against Athena to collect domain names ending in ".edu" in the year 2021.

For gathering ".gov" domains there is an actual government site that list all the currently registered government websites in the United States [8]. The list of .gov domains gathered were stored in a csv file.

3.3. Database Used

We used MYSQL database to maintain a consistent repository for the domain information used in this research. This database stores domain names, SSL, and TLS versions, HTTP status, HTTPS status, and HSTS status to aid in the analysis results section. MYSQL Workbench, a database application, was used to import all the domains that were gathered and stored as csv files in the previous section into the MYSQL database. On an ethical note, to note, the database is a local database housed within the virtual Ubuntu to protect the confidentiality of domain results.

3.4. Scanner for SSL/TLS Identification

We invoked a well-documented tool called SSLSCAN [9] to scan the US-related domains for SSL/TLS protocol versions. This command-line tool accepts a file of domain names as input and returns in XML format the SSL/TLS protocol versions for each domain, if any. The domain names were obtained using a query against the MYSQL database and then converted to csv files to serve as the input for the SSLSCAN tool. Following that, ten instances of the SSLSCAN with the csv files were run to pipeline the scanning process. The total time to scan was thirteen days.

3.5. Python Parsers

To transfer data from SSLSCAN's XML output files to a MYSQL database, a Python application parsing the XML file output was written. An update statement within the script was executed for each domain parsed to keep the database in sync with the SSL/TLS protocol information from the XML files.

To determine whether HSTS was present in the domains collected, another Python script was written to request the domain's header information. This script took as input a csv of domain names. We queried the MYSQL database for HTTPS domain names that had successful scans with the SSLCANNER. As with the previous Python script, this one updated the MYSQL

database in response to the presence of HSTS for each domain. The program execution took five days to process the domains.

This research establishes a clear roadmap by describing the methodology used for domain collection. Furthermore, describing the various tools for domain scanning and storing results provides valuable insight into the research plan's construction.

3.6. Proposed Scoring System

The proposed scoring system is described in this section and takes into account the protocol version and whether HSTS is being used. Additionally, this section presents some examples to properly illustrate the scoring system in practice with given domain configurations.

The scoring system grading is in the numerical range from 0 – 100. We consider a score of a 70 to be passing while anything lower is a failure. The numeric scoring appears to be unhelpful to the computer science professional conducting the analysis, but the number style benchmark is extremely helpful to the end user. The end user would understand that a score of 70 out of 100 is a passing score rather than simply knowing that a domain's TLS 1.2 configuration is acceptable. One assumption made for this scoring system is that if a domain supports multiple HTTPS protocol versions, then the highest HTTPS protocol version is only considered for the domain’s overall score.

The proposed scoring system is split into three tiers. Tier 1 includes domains that support HTTP protocol. These domains automatically receive a score of 0 due to no security being available for the protocol. Tier 2 consists of the HTTPS protocol with the SSL version variations which includes SSL 2.0 and SSL 3.0. SSL 2.0 is assigned a starting score of 5. The reason for using 5 as the starting score is to discourage a good score because SSL 2.0 is deprecated and the oldest HTTPS protocol version. The path for potential updates can be described as seen in Figure 2 where each transition to the next state is awarded 5 points. This pattern continues until the highest SSL protocol version with HSTS is reached which is awarded 20 points. Tier 3 has the same principle as Tier 2 using TLS version variations, but TLS protocol 1.0 starts off with 30 points.

Each transition to the next state for TLS versions is awarded 10 points. The reasoning for using 10 as the even increment across the tier is because TLS is the current generation of secure protocols which should be weighted higher in scoring than Tier 2 is.

Commented [STH8]: Why 10?

Commented [dd9R8]: Answering this question right after the sentence.

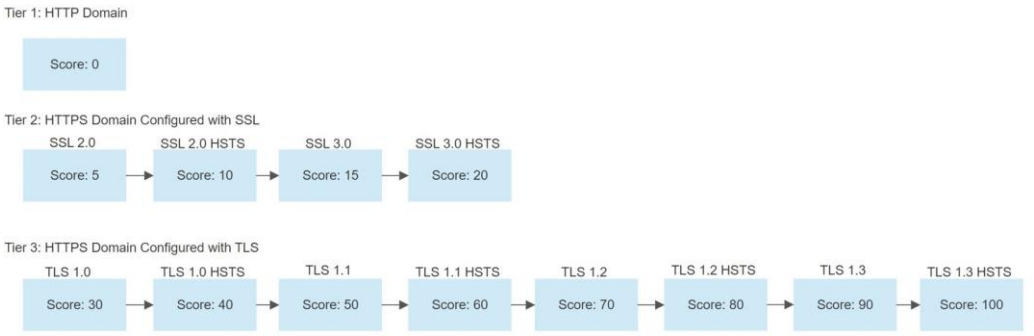

Figure 2. Three tier process for proposed scoring system

A passing score for this scoring system is when TLS 1.2 protocol is used which is awarded 70 points. The reason for this decision is due to RFC officially announcing the deprecation of protocols SSL 2.0, SSL 3.0, TLS 1.0 and TLS 1.1. The best-case scenario is that a domain contains TLS 1.3 HSTS which is a score of 100.

Commented [STH10]: What is the point of the points if you have a set cutoff?

Commented [dd11R10]: Your question is answered in section 3.6 2nd paragraph

We use three examples to illustrate the scoring system. Example one is relatively simple with Figure 3 having the current configuration of just HTTP and because of this we give a failing score of 0. Example two in Figure 4 has SSL 3.0 with HSTS. Since we are in Tier 2 category, we start off with a base score of 5. Since we transitioned three states in order to reach SSL 3.0 with HSTS we add an additional 5 points per state which leads to a total score of 20 points. Example three shows Figure 5 with the current configuration of TLS 1.1 with HSTS. Notice that this figure shows the domain supporting earlier protocol versions. We ignore the earlier protocol versions and only consider the highest. Since we are in the Tier 3 category, we start off with a base score of 30. Since we have transitioned three states in order to reach TLS 1.1 with HSTS, we add an additional 10 points per state leading to a total score of 60 points.

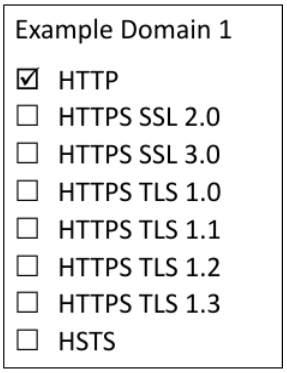

Figure 3. Scoring system example domain with HTTP configuration

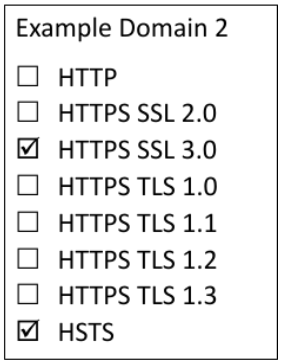

Figure 4. Scoring system example domain with HTTPS SSL 3.0 and HSTS header

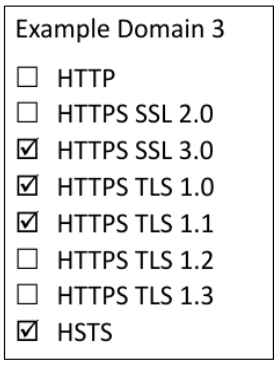

Figure 5. Scoring system example domain with HTTPS SSL 3.0, TLS 1.0, TLS 1.1, and TLS 1.2 and HSTS header.

3.7. HSTS Awareness Survey

During this research, a survey was conducted to determine whether or not people were aware of HSTS as a general subject. "Survey Monkey" was used to track the statistics. A group of about a hundred computer science/IT participants from the Department of Energy's Savannah River Site were polled. The identities of the participants were kept anonymous to protect their privacy. The ultimate goal of using this survey is to introduce the concept that the results may have an impact on HSTS awareness.

4. EXPERIMENT RESULTS/ANALYSIS

4.1. Data Collection

In order to provide a credible basis for the results and analysis, the actual data for this research must be demonstrated. A SQL script file was created to ensure we had a means of recovering the data due to an unforeseen accident resulting in data loss. The files are located in the GitHub repository's "DATABASES" directory. To avoid releasing a domain's important security configuration to the general public, privileged access to the GitHub repository must be acquired as well [17]. Before going over the results of the domains it is necessary to outline some decisions made before and during the analysis. There were originally 1,839,452 domains gathered and transferred to the MYSQL database. 11,723 of the domains did not have the correct extension stemming from the ".us" zone file provided. For example, there was a domain named "100plusus". It was ambiguous if the name should have been "100plusus.us" or "100plus.us". These 11,723 domains were removed from the database to remove this ambiguity. Another consideration made was during the SSLSCANNER application being ran on the domains. There were errors logged in the XML file for each domain that encountered an issue. The same errors were also captured in the python script gathering HSTS header presence. These issues ranged from refused connections from the domains or timeouts. The total number of usable domains after the scanning was 658,500 with 1,169,229 being discarded from the results due to the errors mentioned earlier. Due to the nature of scanning domains, we ensured that the results are only stored in a private repository in GitHub to ensure best ethical practices.

4.2. Overall Domains: HTTP vs HTTPS

Of the 658,500 domains, 38% had only HTTP protocol support while 62% had only HTTPS protocol support shown in Figure 6. Even in modern times, the percentage of US-related domains that are HTTP is alarming. One possible explanation is that the nature of the domains containing the HTTP protocol does not transmit sensitive data over the network at all, implying that HTTPS is unnecessary. This assumption is later disproven when a small random sample of the analysed HTTP related websites were chosen for investigation. We found there were several instances of .edu domains that were HTTP containing login features which is not good practice. To truly determine whether the HTTP vs HTTPs trend is good or bad, a more thorough analysis of all HTTP domains examined in this study is required and is mentioned in the conclusion section of the research.

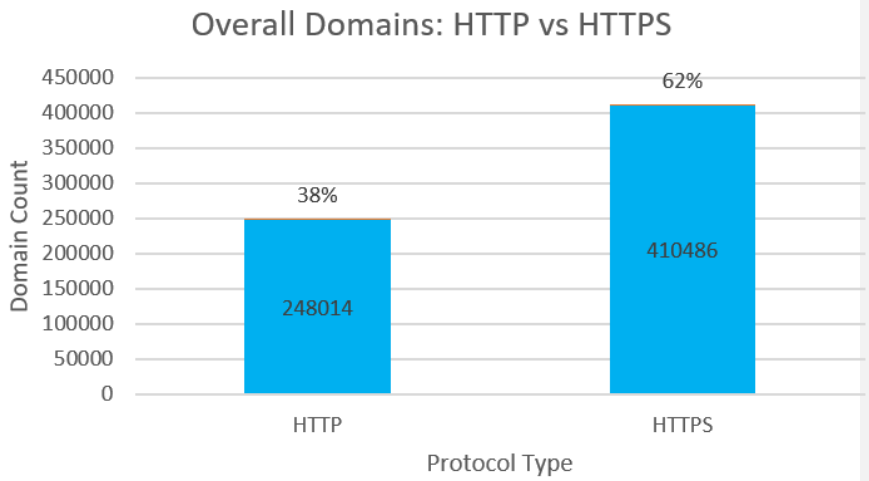

Figure 6. A bar graph visualizing the number of domains scanned that were HTTP protocol or HTTPS protocol.

4.3. Group Specific Domains: HTTP vs HTTPS

We compare domain groupings using the percentage metric. The domain group ".gov" and ".us" were the two domain groups that used the HTTPS protocol the most and the least, respectively.

The domain group ".us" had the largest concentration of HTTP, whilst the domain group ".edu" had the lowest concentration. Since the government is involved and has specialized resources and infrastructure, initially it was believed that the ".gov" domain would have more HTTPS sites, but this was proven false as seen in Figure 7.

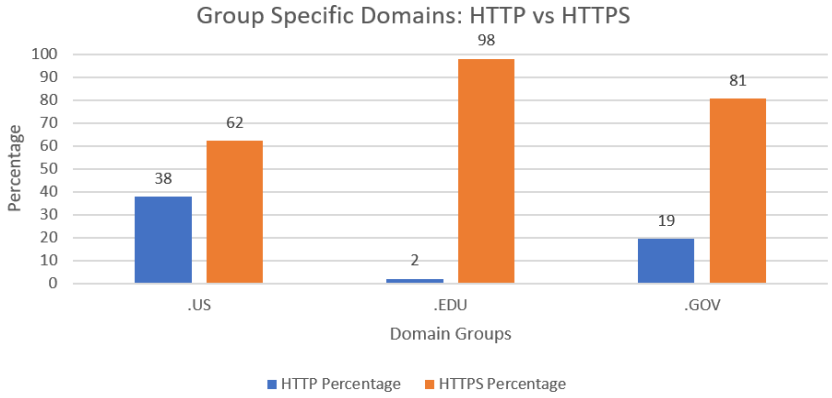

Figure 7. A bar graph visualizing the occurrence of domains by group scanned that were HTTP protocol or HTTPS protocol.

4.4. Overall Domains: HTTPS Protocol Type

We next further split the HTTPS into its individual components SSL 2.0, SSL 3.0, TLS 1.0, TLS 1.1, TLS 1.2, and TLS 1.3 for analysis. It is important to note that domains can have more than one version of HTTPS enabled. Of the 658500 domains analysed, under 1% of the domains were configured with SSL 2.0 and SSL 3.0; 23% of the domains contained TLS 1.0; 24% contained TLS 1.1; 62% contained TLS 1.2; 30% contained TLS 1.3. (See Figure 8)

When analysing the HTTPS protocol versions, the SSL version 2.0 served as the minimum for the number of domains. This met expectations due to it having been deprecated since 2011 by RFC 6176. Additionally, SSL 2.0 was released over two decades ago with vulnerabilities present in them that would create a high need to transition to the TLS protocol. [10]

When analysing the HTTPS protocol versions, TLS 1.2 served as the maximum for the number of domains. This trend is furthermore supported by Qualys SSL Labs. Qualys SSL Lab's past

history of domain scans from January 2021 to October 2021 revealed TLS 1.2 served as the maximum for domain usage [11].

One explanation for why these domains support older protocol versions is that browsers do not enforce the most recent protocol version. TLS 1.2 is now enabled by default in Microsoft Edge, Google Chrome, Apple Safari, and Mozilla Firefox [20]. This situation causes the client-server connection to use the TLS version that is supported by both the browser and the server.

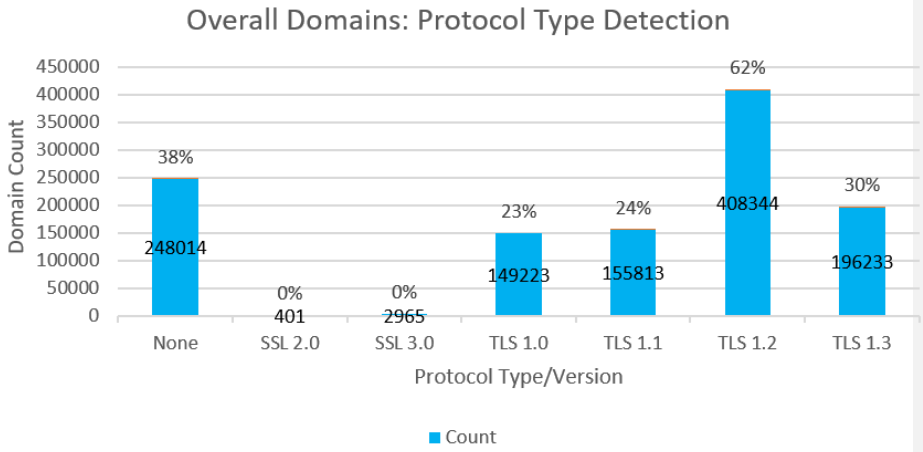

Figure 8. A bar graph visualizing HTTP domain counts along with a more broken-down analysis of HTTPS protocols with their varying versions.

4.5. Group Specific Domains: HTTPS Protocol Type

From a domain group standpoint, the ".edu" domain has the highest occurrence of TLS 1.3 and TLS 1.2, followed by ".gov" and ".us." It is worth noting that ".edu" not only had the highest occurrence of HTTPS domains, but also the highest occurrence of the most recent protocols, as shown in Figure 9.

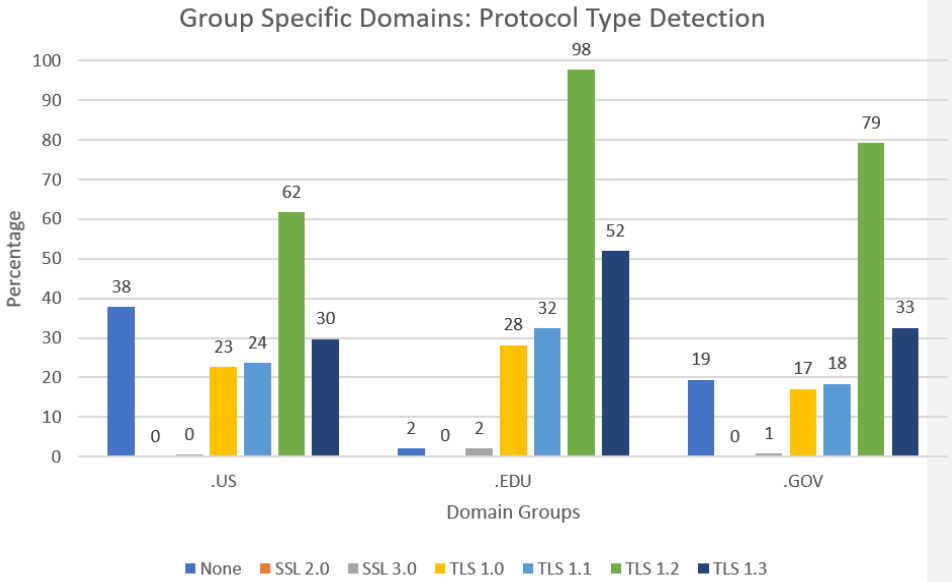

Figure 9. A bar graph visualizing HTTP domain occurrences by domain groups along with a more broken-down analysis of HTTPS protocols with their varying versions.

4.6. HSTS Prescence in Overall Domains

Finally, for HSTS detection of the 658,500 domains analysed, 5% of the domains contained HSTS headers while the other 95% contained no HSTS headers. (See Figure 10). We believe the main culprit for why HSTS headers are low in presence is due to users not being educated or informed about HSTS. More importantly, IT professionals or computer scientists are the community of individuals who would configure the HSTS headers for domains. To explore this theory, a survey was conducted among computer science professionals and IT professionals from a Department of Energy owned facility called Savannah River Nuclear Solutions. The results (see Figure 11) found that 80% of the surveyed individuals did not know what HSTS is.

Commented [STH12]: Could it also be that browsers are not phasing out support for older versions?

Commented [dd13R12]: Good point you brought up during the defense. There could be additional costs from a hardware standpoint that may not warrant the need for browsers to go forward with this. I will put this point towards the HTTPS protocol versions results section (4.4) since browsers automatically enforce the HSTS if it finds it in the server's header response

Commented [STH14]: You completed this survey? If so, why is it just mentioned here?

Commented [dd15R14]: Leaving this sentence here but adding a section within Methods called HSTS Awareness Survey

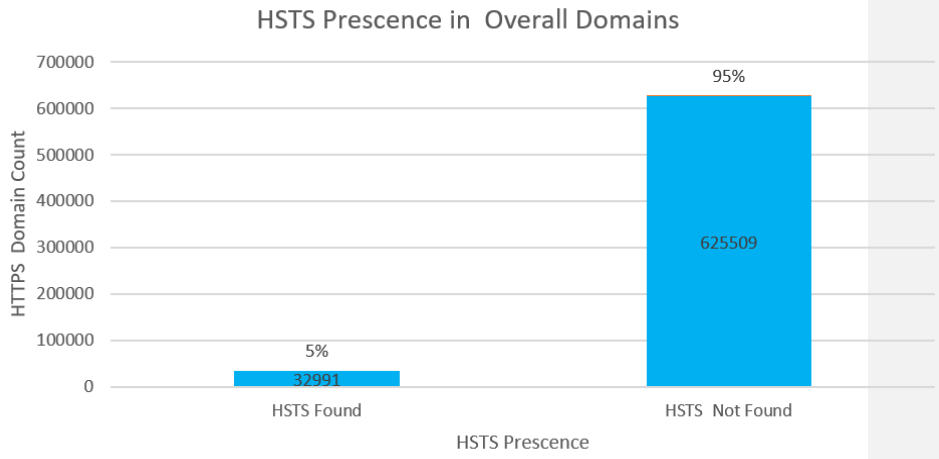

Figure 10. A bar graph visualizing the number of HSTS headers detected versus the number not detected for the scanned domains

Commented [STH16]: Must label axes.

Commented [dd17R16]: Labels are applied now

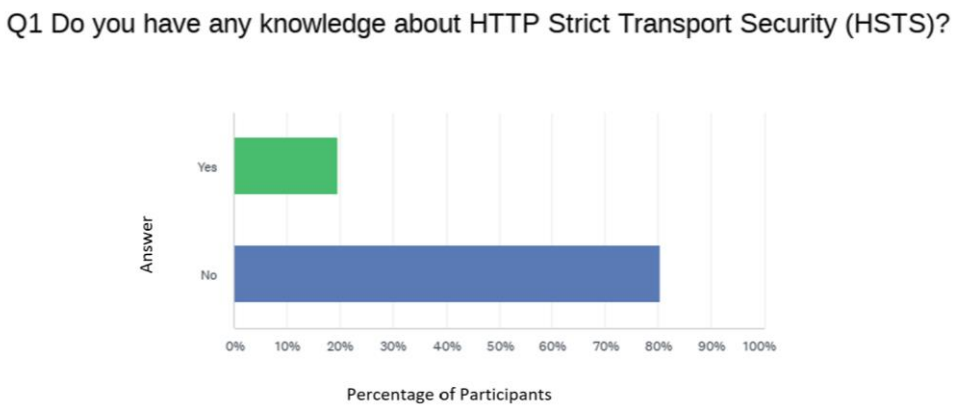

Figure 11. A bar graph from Survey Monkey visualizing the number of participant's knowledge of HSTS.

4.7 HSTS Prescence in Group Specific Domains

Although ".edu" domains had the highest frequency of HTTPS, TLS 1.2, and TLS 1.3, Figure 12 shows that the ".gov" domain had the highest frequency of HSTS header presence. In contrast to the other groups, ".gov" had the highest frequency of domains that did not have HSTS presence.

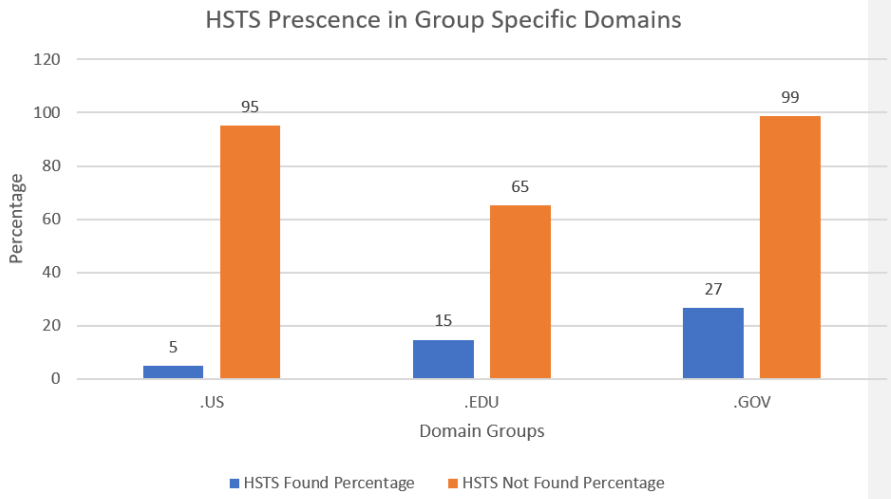

Figure 12. A bar graph visualizing the occurrence of HSTS headers detected versus the occurrence not detected for the scanned domain groups

4.8 Overall Domain: Scoring System Incorporation

We incorporated the proposed scoring system as part of the data analysis for all domains. The results were split into two groups. The first group included domains that contained at minimum TLS 1.2 protocol with or without HSTS configuration and the second group was domains that were below TLS 1.2 protocol with or without HSTS configuration. Figure 13 demonstrates that based on the scoring system rules, 62% of the domains were given a passing score of 70 while 38% percent failed the scoring system. The simple benchmark scoring system can be used as a preliminary report to help establish a focus on acceptable configurations versus unacceptable configurations.

Commented [STH18]: A.k.a., have TLS 1.2 or better? Do not need HSTS?

Commented [dd19R18]: Rephrased sentence to answer question more clearly

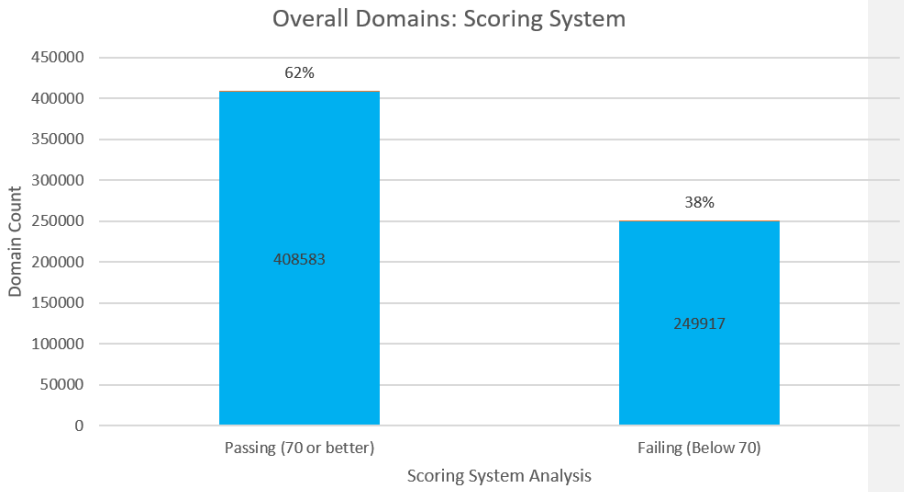

Figure 13. A bar graph visualizing the scoring system applied on the scanned domains.

4.9 Group Specific Domain: Scoring System Incorporation

The same scoring system was used at the domain category level for ".us," ".edu," and ".gov," as shown in Figure 14. Nearly all ".gov" domains met the proposed scoring system's passing criteria, with the highest occurrence, while ".us" had the lowest occurrence of passing scores. The ".us" domains had the highest frequency of failing results from the scoring system, while the ".edu" domains had the lowest frequency of failing results from the scoring system.

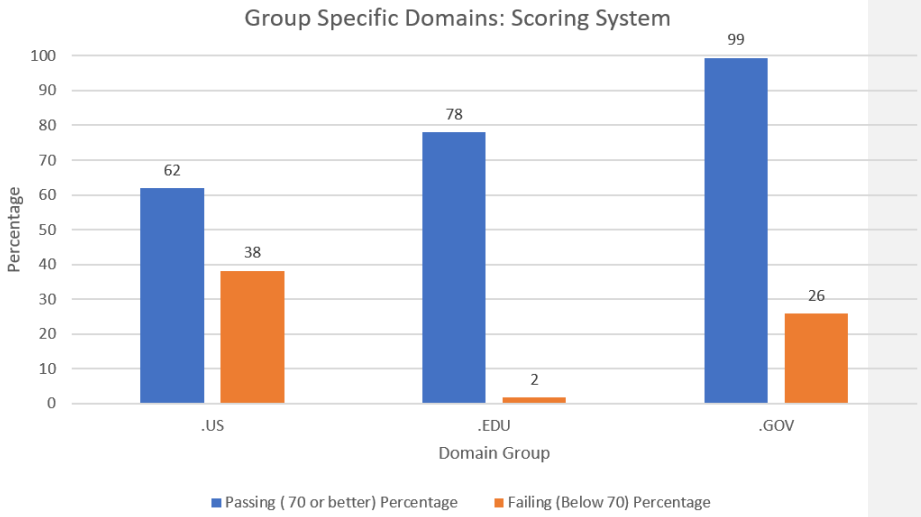

Figure 14. A bar graph visualizing the occurrences of domains with the scoring system applied by domain group.

Overall, there was a vast number of quantifiable results provided in this study. This study, however, had several limitations that were discovered during its development. When it came to analysing the charts from Figures 6-7 in this work, it was difficult to draw any concrete implications due to just having quantifiable information and no other information. For example, if keyword matches such as "Credit Card Number", "Username" and "Password" were employed for HTTP domains, then this would give better indication that these domains should be using HTTPS instead. By having this extra level of detail, then a conclusion could be made about the trend of HTTP vs HTTPS. More thorough research is required to establish the reasoning in the trends identified in the results section. Another limitation in this research was the number of domains gathered. Originally there were over a million domains analysed. Due to network connection denied error logs during the scanning of domains, many were not able to be scanned. In this research we were concerned with only domains that were successfully scanned without issues. Lastly from a coverage standpoint, we lacked ".edu" domains due to not being able to obtain the zone file as explained earlier in the paper.

Commented [STH20]: What non quantifiable information would have been helpful?

Commented [dd21R20]: Keyword matches within webpage input boxes such as "Credit Card Number" , "Username" ,"Password" for HTTP websites. This would indicate that these website may need to be HTTPS instead. This would allow for a conclusion about the trend to be established since we would know the percentage of HTTP domains that shouldn't be HTTP , but instead HTTPS. [ADDING THIS VERBAGE TO THE PAPER]

5. IMPACTS OF RESEARCH

This research has had a wide range of effects. When it comes to ".us," ".edu," and ".gov" domains, there is an established awareness that is known by exposing the current state of network protocol security. Another impact of this research is that the scoring system mechanics could be incorporated in web browsers, allowing end users to determine whether the website's current configuration is good or bad by seeing a numerical score out of a hundred. Most importantly, the surface has been scratched that more HSTS education could be a contributing factor for why there are low occurrences of it in the three domain groups. This opens the door to more research to determine whether education is indeed correlated to HSTS presence in domains.

6. CONCLUSIONS

Commented [STH22]: Before the conclusion, you should have a discussion that describes the impact of your research.

Commented [dd23R22]: Added an Impacts of Research Section

This study has revealed that US-related domains are not up to date with the latest protocol security and HSTS incorporation. By creating a simple scoring system with respect to RFC's most recent deprecations, a general sense of where US based domains stand numerically was easily noticeable. The scoring system rules can be applied to any domain to get a general sense of where their network protocol and HSTS presence stands. This study has also found that HSTS usage in US based domains is low and that the lack of awareness could potentially be one of the contributing factors behind it. It is important to incorporate HSTS knowledge in the workplace for any team that works with configuring network related security which would include education for all stakeholders.

It has also been revealed in the study that a high percentage of US-related domains have enabled lower versions of the HTTPS protocol which needs to be corrected to minimize downgrade related attacks. One path forward would be to notify the end user that owns the domain of this issue, however, due to ethical concerns, the domain names have been kept confidential to ensure privacy. It is therefore more effective to focus on the registrars who sell the domains pre-emptively. EDUCAUSE, GODADDY, and DOTGOV are the companies in this study that handle US-related domain extensions ".edu", ".us", and ".gov" respectively. Another path forward is for

web browsers to enforce the latest TLS version in order to prevent unnecessary downgrade attacks. This is a challenging path forward due to the possibility that clients would have to update their old firmware to support the newer TLS versions.

Future work for this research includes increasing the accuracy of gathering ".edu" domains since the zone file could not be acquired. EDUCAUSE stated that they will only give zone file if they can be positively benefited. If a future researcher partners with other major universities and sends another zone file request to EDUCAUSE, then the chances of them providing the zone file will increase the coverage of ".edu" domains. An additional future work for this research is to perform another assessment of US-related domains in the next coming years and use this research as reference to show if the trend has improved or not. This study aids in the overall understanding of US-related domains and provides a compelling argument that the organizations in charge of facilitating these domains are made aware that there is susceptibility in many of the websites.

ACKNOWLEDGEMENTS

The assistance provided by my professor Dr. Yu-Ju Lin was greatly appreciated. He provided me a successful path of how to organize this thesis and offered valuable input during its development. Additionally, I would like to thank Patrick Hill for allowing me to build upon his research work and utilize in this paper. I would also like to thank Raymond Wilcauskas, a former employee contracted by the Department of Energy who agreed to be part of the reviewing committee for my thesis.

REFERENCES

[1] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, and T. Berners-Lee, "Hypertext Transfer Protocol -- HTTP/1.1," www.rfc-editor.org, Jan. 1997, doi: 10.17487/RFC2068.

[2] R. Oppliger, SSL and Tls: Theory and Practice. Norwood, Ma: Artech House, 2016.

[3] J. Hodges, C. Jackson, and A. Barth, "HTTP Strict Transport Security (HSTS)," Nov. 2012, doi: 10.17487/rfc6797.

[4] "Domain Name Registry Services from GoDaddy Registry," registry.godaddy. https://registry.godaddy/ (accessed Apr. 21, 2022).

Commented [STH24]: Of these sources, I only see two papers that listed as coming from a peer-reviewed conference or journal.

Commented [STH25R24]: They both come from the same journal! You need more peer-reviewed sources to support your work.

Commented [dd26R24]: Source 2 is a book that was peer reviewed by other authors
Source 12 is peer reviewed(Patrick's research)
Source 14,15 peer reviewed (Institution of Research and Technology)
Source 16 was peer reviewed by NDSS conference in 2015
Source 18 is peer reviewed (*Scientific Bulletin of Naval Academy is an academic journal)*
Source 19 is peer reviewed (*INTERNATIONAL CONFERENCE ON SCIENCE AND APPLIED SCIENCE)*
Source 21 is peer reviewed (usenix security conference)

[5] "Hipo/university-domains-list," GitHub, Sep. 02, 2020. https://github.com/Hipo/university-domains-list

[6] "Common Crawl," Common Crawl. https://commoncrawl.org/

[7] "Amazon Athena - Serverless Interactive Query Service - Amazon Web Services," Amazon Web Services, Inc. https://aws.amazon.com/athena/?whats-new-cards.sort-by=item.additionalFields.postDateTime&whats-new-cards.sort-order=desc

[8] "Data | .gov," home.dotgov.gov. https://home.dotgov.gov/data/ (accessed Apr. 21, 2022).

[9] rbsec, "sslscan2," GitHub, Apr. 21, 2022. https://github.com/rbsec/sslscan/ (accessed Apr. 21, 2022).

[10] N. Sullivan, "Why TLS 1.3 isn't in browsers yet," The Cloudflare Blog, Dec. 26, 2017. https://blog.cloudflare.com/why-tls-1-3-isnt-in-browsers-yet/#:~:text=The%20reductive%20answer%20to%20why%20TLS%201.3%20hasn%E2%80%99t

[11] "Qualys SSL Labs - SSL Pulse," www.ssllabs.com. https://www.ssllabs.com/ssl-pulse/

[12] P. Hill and Y.-J. Lin, "Evaluation of Trust Worthiness of State and County Government Websites." Accessed: Apr. 21, 2022. [Online]. Available: http://gator3168.temp.domains/~patriill/wp-content/uploads/2021/05/SAM21-1.pdf

[13] L. Garron, A. Dropbox, and D. Boneh, "The State of HSTS Deployment: A Survey and Common Pitfalls." Accessed: May 13, 2022. [Online]. Available: https://garron.net/crypto/hsts/hsts-2013.pdf

[14] W. J. Buchanan, S. Helme, and A. Woodward, "Analysis of the adoption of security headers in HTTP," IET Information Security, vol. 12, no. 2, pp. 118–126, Mar. 2018, doi: 10.1049/iet-ifs.2016.0621.

[15] S. De los Santos and J. Torres, "Analysing HSTS and HPKP implementation in both browsers and servers," IET Information Security, vol. 12, no. 4, pp. 275–284, Jul. 2018, doi: 10.1049/iet-ifs.2017.0030.

[16] M. Kranch and J. Bonneau, "Upgrading HTTPS in mid-air: An Empirical Study of Strict Transport Security and Key Pinning," Proceedings 2015 Network and Distributed System Security Symposium, 2015, doi: 10.14722/ndss.2015.23162.

[17] De'Jean Dunbar, "thesis," GitHub, Sep. 16, 2022. https://github.com/dunbarmustard/thesis (accessed Sep. 16, 2022).

[18] D. Glăvan, "Man in the middle attack on HTTPS protocol," *Scientific Bulletin of Naval Academy*, vol. XXIII, no. 1, pp. 199–201, Jul. 2020, doi: 10.21279/1454-864x-20-i1-026.

[19] A. Purwanto and A. W. R. Emanuel, "The state of website security response headers in Indonesia banking," *INTERNATIONAL CONFERENCE ON SCIENCE AND APPLIED SCIENCE (ICSAS2020)*, 2020, doi: 10.1063/5.0030359.

[20] "Enable Support for TLS 1.2 or 1.3 on Web Browsers," *www.technology.pitt.edu*, May 19, 2021. https://www.technology.pitt.edu/help-desk/how-to-documents/enable-support-tls-12-or-13-web-browsers (accessed Nov. 16, 2022).

[21] A. P. Felt, R. Barnes, A. King, C. Palmer, C. Bentzel, and P. Tabriz, “Measuring {HTTPS} Adoption on the Web,” *www.usenix.org*, 2017. https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/felt

Authors

DeJean Dunbar earned his B.S in Computer Science from Charleston Southern University in 2017. He is currently pursing his masters degree at Charleston Southern University. DeJean’s interests in computer science includes automation-based systems, database management systems and networking security.