Analysis of Domain Connections Across Websites

Nick Thompson, Giselle Briand, Colin Groh, Alon Neerman, Matthew O'Donnell

CY4740 Network Security – December 18, 2020

Purpose and Scope of Project:

We chose to pursue an audit of the tracking and information sharing by a number of top sites as ranked by the Tranco list[1] on Dec. 16, 2020, across social media, news sites, e-commerce, and other highly trafficked entities. We audited these sites through the use of *mitmproxy* (a Python reverse proxy tool), which allowed us to capture all of the domain connections made during a site visit and gain insights on the number and pervasive nature of some trackers. Using these captured requests, gathered through a custom developed *mitmproxy* addon, we were able to write scripts to parse the dumped data into readable data we could analyze. Through this information, we were able to examine the depth of third party trackers across websites as well as their current security practices.

From our initial proposal, which was a little more nebulous and loosely defined, we decided to focus on the latent tracking of a consumer's general browsing behavior (see Methodology for more detail). Instead of using a large set of domains from the Tranco list, which is composed of over one million domains, and crawling those public paths, we chose to manually browse a set of up to 62 domains. Using a smaller subset allowed us to evaluate traffic, the connections between different domains, and communications to ad networks with heavier detail. On each domain, we mirrored human behavior (clicking, hovering over ads, creating accounts, etc.) to pull more realistic information about what domains might be contacted during a session and how many times.

While our original scope involved a more automated approach of creating a Chrome Extension to automatically tag domains visited, we decided to implement the tagging functionality manually for each time a new domain is visited. This was done to reduce the scope of the project in the interest of time. Similarly, we chose to analyze a smaller subset of domains in order to provide a more detailed picture of what each site is connecting to. Manual tagging would also allow any device to use the proxy, instead of only desktop web browsers that support

[1] https://tranco-list.eu/list/9JL2

extensions. In the future, our tool could be used on things like smartphones and IoT devices to see if there is any difference in traffic behavior.

Methodology:

Using an installation of Firefox, we disabled all tracking protection, as well as the Online Certificate Status Protocol (OCSP) and Telemetry to Firefox. We did this to reduce false positives of requests that were coming from the browser itself, and not the current website we were on. We then used *mitmproxy* to create a proxy server to capture a visit to a single site, which we manually tagged before spending roughly a minute on site exploration and ad impressions. Once data collection was complete, data was dumped via JSON and then parsed using the *pyvis* Python library. This allowed us to create a graph of domain connectivity, with the most commonly shared trackers in the center of the visualization. We also had a couple other scripts to collect statistics on the usage of TLS, information about the certificates used, and the count of all domains in a dump.

Results:

The result of our proxy and data collection is represented by deliverables including four interactive graphs, each of which is accompanied by three statistical text files offering different analyses. The interactive graphs represent visited nodes (chosen domains) from that group, along with public domains connected to each node that were also visited by our proxy server. Certificate information, domain connectivity, and TLS connections were detected and collected across all visited domains. Certificate issuers and TLS connections were specifically chosen for analysis in this project to evaluate the security strength of the domains visited and to determine how many "unsafe" domains there are among the ones that were tracked. Furthermore, we also wanted to study the difference between tracked domains in an authenticated state versus an unauthenticated state to evaluate the difference in reach between connected domains when a user is logged in/out.

The four graphs represent four distinct domain categories we chose to pursue: (1) news outlet sites, (2) top websites from Tranco's list (biggest group), (3) top China-based sites, and (4) authentication-based sites. We found these groups to be relevant, yet different enough as subjects for ad network traffic evaluation. News outlets harbor an exorbitant amount of external source links and vast advertisement content as well, making it a goldmine-like group to study, knowing

it would produce a lot of data. We chose to also use a big dump of popular domains from Tranco's list because a larger sample would allow us to really extend the scope of this project and see how far-reaching domain connections can go with more nodes in question. The implications of this are even heavier, since the nodes being tested are some of the most popular and frequently visited domains on the Internet. We also decided to collect domain data from China-based domains to compare their security and connectivity to our other data. We believed this would be an interesting group to analyze, given China's technological infrastructure and its extreme isolation and restrictions from the rest of the Internet. And, finally, we believed authentication/no-authentication data would be fascinating to analyze, as it would showcase the different depths of domain traveling between ad networks when a user is logged in versus when they are not.

For each group, there were particular aspects of the domains we were interested in analyzing: TLS data, domain connection data, and certificate issuer data, all together manifesting into the 3 statistical text files for each graph/domain group. The TLS statistics provide insight on which TLS versions are in use for that node along with the proportion of its connected domains using that TLS version and how many aren't using TLS at all (if applicable). The domain data lists all connected domains and ad networks for each starting node and the number of times they're accessed on the site. We also wanted to analyze the certificate lengths in the domains we traversed, given the new regulation issued by browsers that all SSL/TLS certificates cannot have a lifespan longer than 13 months, or 397 days.[2] Certificate issuer data lends a deeper look into the security of the nodes by examining who issues certificates and if any are expired/have not yet abided by the new browser regulations.

News Dump

Starting with the data collected by the news dump, a subset of 28 news source domains and their connected domains were tracked, with a total of 54,862 domains analyzed. It was found that across all of them, 32,277 use TLS 1.3 (59%), while ~22,500 use TLS 1.2 (41%). Luckily, only 96 domains (0.2%) use no TLS at all. Among the 28 nodes, Wired had the most connected domains-- 4,130, with CNN just behind at 3,800. The top-used/connected to domains across the nodes that were visited were Google's ad service platform and Double Click (also owned by

[2] https://www.globalsign.com/en/blog/maximum-ssltls-certificate-validity-now-one-year

Google), both of which were connected to all 28 original domains. This makes Google the "worst offender" for news sites' ad traffic from our data pool.

For certificate issuers, DigiCert is the most commonly used CA, with over 40% of visited domains using their service. The second top contender is Google Trust Services, in use by just over 20% of domains. No certificates in use by any analyzed domain is expired, but there are several with lifespans exceeding the 13-month limit. CloudFare, Amazon, Google Trust Services, and Let's Encrypt are among some of the only issuers that have all certificates within the 397-day lifespan.

Big Dump

In the bigger dump, 62 nodes were traversed (50 in addition to the original 12) to broaden the scope of the domain network. A total of 33,070 domains were reached and analyzed in this dump. Twice the amount of nodes were included in this run from the news dump, yet it yielded far less domain connections (this is a reasonable conclusion, given news sites tend to be more populated with hyperlinks, ads, and external sources). In this dump, TLS 1.3 is also used most frequently, with just under 65% of domains. TLS 1.2, alternatively, is used by roughly 35% of traversed domains. Oddly, even fewer domains use no TLS in this dump than in the previous, smaller dump; only 113 domains, or .34%. 3 nodes have 100% of their connected domains using TLS 1.3 (Mozilla, Instagram, and Wikipedia)-- all others have a combination of 1.3, 1.2, and no TLS. Fox News resulted in being the domain with the highest number of domain connections in this dump at 2,374, although AOL was just behind at 2,337.

The most commonly connected domains were also captured. At the top of the list is www.google.com, which was cross-used over 55 different domains (almost every original node), with www.adservice.google.com, Google's ad service platform, as the second highest across 38 domains. This signifies that the leading connective domain is Google and its ad service across the nodes and domains that were studied in this group. As far as certificate issuers, overall, DigiCert was the top issuer across all domains, with over 15,000 domains (46%) using their validation services. Other popular certificate authorities are Google Trust Services, Amazon, GlobalSign, and Let's Encrypt (the only CA service in this dump's top 5 that is free). Certificate lengths for this dump are extremely variable. Despite the industry's new regulations for shorter certificate lifespans, there are still thousands of certificates in use that have lifespans beyond the newly-established 397 days, some upwards of over 1,100 days until expiration. While that may

not imply an immediate security risk, long-lasting certificates may have other security implications, especially if they're not abiding by industry standards. These longer certificates are grandfathered in to being valid, with only certs created after this past September 1st having to abide by the new 13 month rule. Once again, there were luckily no expired certificates found in the dump.

China Dump

In the third dump, we focused on 14 China-based top domains to traverse through (360, Baidu, JD, QQ, Sohu, Weibo, Zhanqi, Tianya, Xinhuanet, Tmall, Sina, AliPay, AliExpress, and CSDN). The total number of domains visited through these initial nodes is 6,236. Of these domains, 3,800 (61%) of them use TLS 1.2, while 1,562 (25%) use TLS 1.3. Meanwhile, over 14% of domains use no TLS whatsoever. This is clearly strikingly different from both previous dumps collected. Seeming strange, we did some research and it was discovered that the Great Firewall (GFW), China's efforts to strictly prohibit Internet traffic, access to foreign websites, and force censorship, blocks all traffic that uses TLS 1.3.[3] TLS 1.3 is considered to be safer and faster than its predecessor TLS 1.2. Even though the sample size of this China-based domain dump isn't as big as our biggest dump, having the majority of domains using TLS 1.2 over TLS 1.3 while also potentially rejecting any TLS 1.3 connection could pose a security risk. It also shows how different security infrastructure and domain connectivity are for these Chinese domains compared to the U.S. CSDN garnered the most number of connected domains, 1,073, and was also the node that had the highest percentage of connected domains using TLS 1.2 (over 87%).

The domain with the highest number of cross-usages over the different domains accessed was www.google.com. Across all Chinese domains, DigiCert was once again the most commonly used issuer, with 52%. Interestingly, over 14% of the visited domains used no TLS at all, ergo had no CA usage. Similar to the bigger dump, there were no expired certificates among the ones that were detected.

Auth/No-Auth Dump

The final dump related authenticated behavior on a domain with unauthenticated behavior on the same domain. 3 nodes with user account capabilities were chosen: Reddit,

[3] https://www.zdnet.com/article/china-is-now-blocking-all-encrypted-https-traffic-using-tls-1-3-and-esni/

Amazon, and YouTube. A total of 3,308 domains were reached across all auth and non-auth instances. TLS 1.3 was again the most popular with 65% of domains. There are no drastic differences of TLS usages between authenticated and unauthenticated instances. Interestingly, however, YouTube uses a combination of TLS 1.2, 1.3, and no TLS when a user is logged in, but exclusively (100% of the time) uses TLS 1.3 when a user is *not* logged in. In terms of connected domains, on average, authenticated sites send out significantly more requests to ad networks than unauthenticated sites. For example, both Reddit and YouTube sent twice as many requests when a user was logged in versus when the user was not. While the sample is small, the results are very telling of domain and ad network behavior for authenticated site visitors, which evidently facilitates far more traffic between advertisement domains.

For certificate issuers, DigiCert was the leading CA for 59% of domains accessed. On average, more CAs were used in the authenticated sites than the unauthenticated ones. DigiCert leads Reddit and Amazon, while YouTube predominantly utilizes Google Trust Services-- 93% of the time for authenticated users, and 100% of the time for unauthenticated users.

Conclusions:

To overview the security hygiene of the visited domains in this study, no domain across any of the four studied groups utilizes TLS 1.0 or TLS 1.1, which are both considered to no longer be secure. Similarly, there were no expired certificates found, which would also, of course, be a security concern. Among the certificate issuers seen in our collected data, there were no known “unsafe” CAs in use, even though several hundred domains were using certificates that had life spans longer than three years. Although we didn’t necessarily expect unsafe certificate issuers or TLS versions to be used by domains on Tranco’s list, some of their connected domains may have which could become vulnerabilities for the parent node. Some false positives may have been captured due to ongoing connections that had yet to finish when a new capture was started, or requests from Firefox itself. We did our best to avoid this by closing all existing tabs before beginning a new capture of a different site, but there may have been hanging connections that we couldn’t see. We disabled all Firefox telemetry and security settings, but there could have been some built-in behavior that we missed or couldn’t be blocked (like captive portal or software update checks).

Throughout this project, we were able to monitor all background web traffic that stemmed from normal behavior on different sites. After collecting our data through our

man-in-the-middle proxy we were able to visualize all shared connections between seemingly unrelated domains. While our total sample sizes were small, relatively speaking, our findings were telling and evidential about what we as a group knew coming into this project (and what this class reinforced): everything is always more connected than we think. Collecting data sets and visualizing that connectivity provided a greater insight into the scope of our digital footprint and the power of one click. And while this data doesn't represent all of Internet traffic through these sites, it gives a glimpse of just how far our data travels, even when we think it's not.

The final source code, data dumps, and analysis for the project is available on Github at https://github.com/inickt/cy4740-final-project.