Security Analysis of the Chinese Web: How well is it protected?

Ping Chen[†], Nick Nikiforakis[‡], Lieven Desmet[†], Christophe Huygens[†]

[†]iMinds-DistriNet, KU Leuven, 3001 Leuven, Belgium
[†]{firstname.lastname}@cs.kuleuven.be

[‡]Department of Computer Science, Stony Brook University
[‡]nick@cs.stonybrook.edu

ABSTRACT

As the web rapidly expands and gets integrated into the daily lives of more and more people, so does the number of cyber attacks against it. To defend against attackers, website operators can utilize a wide range of defense mechanisms, both at the server-side, as well as the client-side of their web applications. From a security-metrics standpoint, the presence or absence of these mechanisms can be used as a security indicator of any given website.

In this paper, through a large-scale analysis of the 10,000 most popular Chinese websites, we analyze the security of the Chinese web by investigating the usage of client-side security policies, and evaluating the discovered HTTPS implementations. We show that, when compared to popular websites of the rest of the world, a significant fraction of Chinese websites lag behind on the adoption of good security practices. Among other findings, we report on the fact that 6% of websites inadvertently leak private user information, such as Chinese identity numbers, by placing spreadsheet files with sensitive content in directories indexed by search engines.

1. INTRODUCTION

The web becomes more and more popular and important in China, with more than 1.47 million websites on the Chinese web and 44% Chinese people (591 million) using it as of July 2013 [17]. While Chinese Internet users enjoy the convenience and flexibility that the web brings them, and Chinese companies heavily depend on the web for their business operations, at the same time, the web also draws increasing attention from attackers.

In December 2011, Chinese hackers leaked six million user accounts from `csdn.net` (website of Chinese software developer network), and 40 million user accounts from `tianya.cn` (the largest Chinese online forum) [5]. Over 20 million hotel reservation records containing customers' private information were leaked on the Chinese web in December 2013 [1], and more than seven thousand vulnerabilities found in Chinese websites were reported in 2013 [12]. These incidents raise concerns about the security of Chinese web, and show that even high-profile websites such as `csdn.net` can be compromised.

As people depend more and more on the web for their daily lives and businesses, it may be desirable for government and supervisory organizations to continuously assess and monitor the security of the web environment. Such an assessment typically involves a large number of websites belonging to a country, or a specific industry sector, and hence it has to be done externally for efficiency, since traditional internal penetration testing and code reviewing for each website is time and labor consuming. Recently, Van Goethem et al. [30] conducted a security assessment for more than 22,000 European websites, and proposed a score system to compare different websites' security levels, showing that such a large-scale security analysis of the web is achievable, albeit challenging.

In this paper, we apply the same basic methodology to investigate the security of the Chinese web through a large-scale experiment. Instead of searching for vulnerabilities and weaknesses, we seek to discover the usage of defense mechanisms by Chinese websites, trying to answer the question "How well is the Chinese web protected?" In particular, we focus on client-side security policies, and HTTPS implementations, both of which can be passively detected. These mechanisms are developed by the security community for securing the web, thus the presence of these mechanisms on a website can be used as an indicator of the security awareness and practices of that website.

Our main contributions are the following: (1) We report the usage of client-side security policies in the top 10,000 Chinese websites, and compare it to the statistics obtained from non-Chinese websites, showing that the Chinese web lags behind with respect to the adoption of client-side security policies; (2) We provide a comprehensive evaluation of HTTPS implementations on the Chinese web, illustrating that the majority of HTTPS adopters do not have secure and protected HTTPS implementations; (3) We present a case study on the inadvertent private data leakage, showing that 6% of websites leak Chinese identity numbers, that are collected in spreadsheet files and can be obtained by search engines.

SafeConfig'14, November 3, 2014, Scottsdale, Arizona, USA.

http://dx.doi.org/10.1145/2665936.2665938.

2. DATA COLLECTION

2.1 Website selection

According to the status report published by China Internet Network Information Center (CNNIC) in 2013, there are 1.47 million websites on the Chinese web [17]. Since checking the whole content of all websites on the Chinese web is close to infeasible due to its large size and registration walls, we focus on the high-profile websites that are ranked in Alexa's list of most popular websites.

Starting with Alexa's list of top 1 million sites, we first select the set of `.cn` domains from the list. Next, we cross-check `top.chinaz.com`, a website providing a ranked list consisting of more than 6,000 Chinese websites, in order to identify the set of Chinese websites not using `.cn` domain (e.g., `baidu.com`, `qq.com`) in Alexa's list. As a result, we obtain a set of more than 12,000 Chinese websites.

We then filter out the websites without an ICP (Internet Content Provider) license. The ICP license is a permit issued by the Chinese government to permit China-based websites to operate. Websites operating in China without an ICP license will be fined or shut down as specified by Chinese Internet regulations. By removing websites without an ICP license from our dataset, we try to avoid the inclusion of some malicious websites involved in illegal online activities, such as phishing and porn websites. After this filtering step, we obtained about 10,000 websites as our targets, to represent the high-profile part of the Chinese web.

2.2 Crawling experiment

In order to assess the security of the top 10,000 Chinese websites, we conducted a crawling experiment to visit the popular webpages from these websites. For each website, we obtain up to 200 webpage URLs by using the Bing Search API [4] with parameters `site:domain` and `Market:zh-cn`. For instance, the search for `site:baidu.com Market:zh-cn` in Bing will return a set of Chinese webpages belonging to `baidu.com`. By setting the parameter `Market:zh-cn`, we instruct Bing to only search for Simplified Chinese webpages originating from Mainland China, while excluding any English or Traditional Chinese webpages that are not targeting users in Mainland China.

After the webpage URLs are obtained, we use a crawler built on top of HtmlUnit [6], i.e., a headless scriptable browser, to visit each URL. By loading every webpage within HtmlUnit and with an appropriately set user-agent, we mimicked the behavior of a regular user visiting a website with an Internet Explorer browser, which is the typical case in China. In total, we analyzed more than 1.4 million webpages for the top 10,000 Chinese websites, with an average of 147 webpages per website.

For a comparison analysis, we collected a set of 10,000 non-Chinese websites, and launched the same crawling experiment for these non-Chinese websites. The set of non-Chinese websites was collected as follows: for each Chinese website included in our study, we randomly select a non-Chinese website with the closest rank from Alexa's top 1 million sites. For example, `yahoo.com` (rank 4) is selected as it is the closest neighbor of `baidu.com` (rank 5).

3. USAGE OF CLIENT-SIDE SECURITY POLICIES

Client-side security policies are declarative security mechanisms, whereby a website communicates its intent and leaves it up to the browsers to enforce it. There are a number of benefits to the use of client-side security policies [19], with auditability being one of them. Client-side security policies are typically sent via HTTP response headers, thus it is straightforward to determine a website's security expectations by passive analysis.

By instructing browsers to enforce protection through server-provided policies, websites can address a number of security issues in a very straightforward manner. Today, several client-side security policies exist, all of which address various security issues such as cross-site scripting (XSS) and clickjacking attacks. Although these policies are not security panacea, their usage on a website can indicate the "security consciousness" of that website and its security objectives.

3.1 Client-side security policies

HTTP-only Cookies First introduced in Internet Explorer (IE) 6 SP1, the `HttpOnly` attribute is designed to mitigate the risk of malicious client-side scripts accessing sensitive cookie values. Cookies are accessible to JavaScript code by default, which allows attackers to steal the cookies via an XSS attack. Using the `HttpOnly` attribute in a `Set-Cookie` header restrict the access of that cookie to the HTTP(S) protocol, making it inaccessible to client-side JavaScript [14].

Content Security Policy (CSP) CSP provides a standard HTTP response header (`Content-Security-Policy`) that allows a webpage to declare approved sources of content that browsers should be allowed to load on that specific page. Whenever a requested resource originates from a source that is not defined in the CSP, it will simply not be loaded [28]. For example, if the policy does not allow in-line JavaScript, then, even if an attacker is able to inject malicious JavaScript in the webpage, the injected code will not be executed. CSP is mainly designed to mitigate data injection vulnerabilities, although it can also help mitigating mixed-content attacks [15]. CSP is a W3C candidate as of 2012 [29], and almost all modern browsers support it.

X-Frame-Options The HTTP response header `X-Frame-Options` is designed to mitigate Clickjacking attacks [26]. In a Clickjacking attack, the attacker redresses the user interface of website A with transparent layers, and then trick the user into clicking on a button on an embed page from website B when they were intending to click on the the same place of the overlaying page from website A. To stop Clickjacking attacks, the `X-Frame-Options` header can be used to instruct a user's browser whether a certain page is allowed to be embedded in a frame. For example, if the `X-Frame-Options` header's value is `DENY`, then the browser will prevent the page from rendering when embedded within a frame.

X-Content-Type-Options Microsoft's IE has a MIME-sniffing feature that will attempt to determine the content-type for each downloaded resource. This feature, however, can lead to security problems for servers hosting untrusted content, since attackers can craft malicious files abusing the sniffing feature in IE. To prevent IE from MIME-sniffing, thus reducing exposure to attacks, a web server can send the `X-Content-Type-Options` header with the `nosniff` value.

3.2 Findings and discussion

Table 1 gives an overview of the usage of client-side security policies on the Chinese web. HTTP-only cookies are much more widely used than other policies, and their usage on Chinese websites is greater than that of non-Chinese websites. The `X-Frame-Options` and `X-Content-Type-Options` header are much less popular, and their usage on the Chinese web is lower than the estimated global statistics. For CSP, the adoption both on the Chinese web and globally is quite low, most likely due to the relative newness of the mechanism.

Since client-side security policies rely on a browser's enforcement, it is important to supply the browser with a correct policy. In our experiment, we found 5 Chinese websites using the `X-Frame-Options` header with an incorrect value, e.g., `SAMEORIGIN/DENY`, which is effectively ignored by the browser. Overall, the adoption of client-side security policies on the Chinese web is low, which indicates that a lot Chinese websites lag behind with respect to adopting countermeasures on the client side.

By analyzing the rank of websites using client-side security policies, we found that there is no strong correlation between a website's popularity and its adoption of client-side security policies, as shown in Figure 1. The figure also shows the distribution of Chinese websites with HTTPS implemented, which will be discussed in next section.

4. SECURITY OF HTTPS IMPLEMENTATIONS

The HTTPS protocol is the standard solution for securing web traffic, which guarantees the confidentiality and integrity of web communications by adding the security capabilities of SSL/TLS to HTTP. It also provides website authenticity with the CA/B (Certificate Authority/Browser) trust model. While HTTPS is designed to provide strong security, it may fail to achieve the desired security goals if it is implemented in the wrong way. A range of security issues associated with HTTPS have, over the time, been discovered, ranging from cryptographic weaknesses and design flaws in SSL/TLS protocol, to the insecure design of HTTPS websites, and bad coding practices [16].

When migrating to HTTPS, websites should try to avoid known security issues, and consider to add extra defenses to HTTPS. In this section, we assess a website's HTTPS implementation in two ways: (1) the presence of known security issues related to HTTPS; (2) the usage of extra client-side security policies for the better enforcement of HTTPS.

Note that when discussing HTTPS security, we use the active network attacker model. An active network attacker positions himself on a network between the host running the web browser and the web server, and is able to intercept and tamper with the network traffic passing by. The attacker can read, modify, delete, and inject HTTP requests and responses, but he is typically not able to decipher any encrypted information.

4.1 HTTPS security issues

To build a secure HTTPS website, there are a number of security pitfalls that websites should try to avoid. In this section, we check whether an HTTPS website suffers from the following security issues: **Insecure SSL/TLS Implementation** As the core part of HTTPS, secure SSL/TLS implementation is critical to HTTPS websites. In our assessment, we use an SSL scanner called sslyze [9] to search for the following security issues related to SSL/TLS: broken certificate validation chain (e.g. untrusted CA), support of insecure SSL 2.0, use of weak ciphers (e.g. export-grade ciphers with small encryption key length), and the vulnerability to insecure renegotiation attacks [11], and CRIME attacks [25]. This assessment is similar to Qualys' SSL survey for the global SSL landscape. According to Qualys' data in March 2014, more than 75% the tested HTTPS websites suffer from insecure SSL/TLS implementations [8].

Post-to-HTTPS Forms It is a relatively common practice in many HTTPS websites to provide a form (such as a login box) on an HTTP page while arranging for any sensitive information to be submitted over HTTPS. This, however, is a bad practice, since an active network attacker can launch an SSL-stripping attack to steal a user's sensitive data without raising the user's suspicion [23].

Mixed-content Inclusion When migrating to HTTPS, many websites fail to fully update their applications, resulting in mixed-content inclusion, which can render the HTTPS protection useless [15]. Mixed-content inclusion occurs when the main webpage is sent over a secure HTTPS channel, while some additional content included on that page, such as images and scripts, are delivered over non-secured HTTP connections. As a result, an active network attacker can still try to compromise an HTTPS website by intercepting and modifying the unencrypted content.

4.2 Client-side security policies for HTTPS websites

In addition to the client-side security policies discussed in Section 3, HTTPS websites can also make use of the HTTP `Strict-Transport-Security` policy and the `Secure` attribute of cookies. These mechanisms are specifically designed for HTTPS websites and can be used to mitigate some HTTPS-specific security issues.

HTTP Strict-Transport-Security (HSTS) HSTS is designed to mainly prevent SSL-stripping attacks where a secure HTTPS connection is downgraded to a plain HTTP connection by the attacker. Set by a website via a HTTP response header field (`Strict-Transport-Security`), HSTS specifies a period of time during which the user's browser is instructed that all requests to that website need to be sent over HTTPS, regardless of what a user requests. The HSTS Policy helps protecting website users against both passive eavesdropping, as well as active Man-in-the-Middle (MITM) attacks.

Secure Cookies HTTPS websites should set the `Secure` attribute when sending cookies to a user's browser, which can prevent cookies from being intercepted by an active network attacker. Although the traffic between a web server and a browser is encrypted when using HTTPS, the cookies stored in the browser are not, by default, limited to an HTTPS context. Thus an active network attacker can intercept any outbound HTTP request from the browser and redirect that request to the same website over HTTP in order to reveal the cookies [14]. By setting the `Secure` attribute, the scope of a cookie is limited to secure channels, thus stopping browsers from sending cookies over unencrypted HTTP requests.

Security mechanisms	% of Non-Chinese websites	% of Chinese websites	Example findings
HTTP-only Cookies	25.7%	43%	`JSESSIONID=+8OuWN7rO7G nE3Ag$dXHpvH8h4rVHLauth ; HttpOnly` (alipay.com)
Content Security Policy	0.06%	0.01%	`script-src *.zhihu.com *.google-analytics.com 'unsafe-eval'` (zhihu.com)
X-Frame-Options	5.8%	1.2%	`DENY` (weibo.com)
X-Content-Type-Options	4.6%	0.5%	`nosniff` (alibaba.com)

Table 1: Usage of client-side security policies on the Chinese web

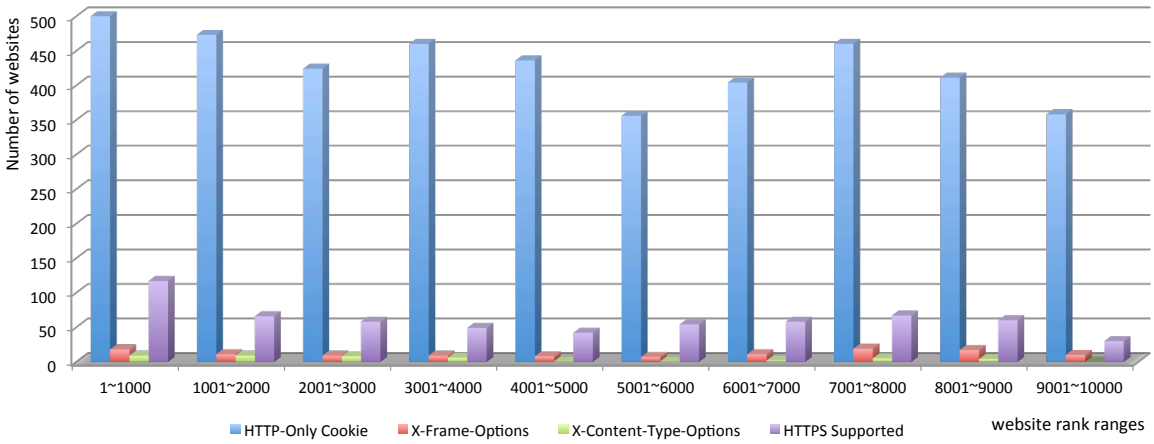

Figure 1: Distribution of websites using client-side security policies over rank ranges

4.3 Findings and discussion

In order to identify HTTPS pages from the 10,000 websites, we try to enumerate HTTPS URLs for each website. First, we select HTTPS URLs obtained from Bing (as described in Section 2). Additionally, we search for any HTTPS links on HTTP webpages during our crawling experiment, and add these HTTPS links to our dataset for later crawling. By doing so, we identified 672 Chinese HTTPS websites, with an average number of 15 HTTPS pages per site. For the dataset of non-Chinese websites, we found 1,601 HTTPS websites, with an average number of 19 HTTPS pages per site. We summarize our findings concerning HTTPS security issues and defense mechanisms in Table 2.

The vast majority (84.1%) of Chinese HTTPS websites have SSL/TLS implementation issues. More specifically, 13% of websites are using self-signed certificates, 19% have insecure SSL 2.0 enabled, 30% support weak ciphers, and 18% and 33% are vulnerable to CRIME and SSL Renegotiation attacks, respectively.

One can see that Chinese HTTPS websites tend to have less problems with respect to post-to-HTTPS forms and mixed-content inclusion, when comparing to the non-Chinese HTTPS websites. Note, however, that we can not claim that Chinese HTTPS websites are doing better than non-Chinese HTTPS websites. Since 28% of Chinese HTTPS websites have only one HTTPS webpage, the probability of these issues occurring in Chinese websites is smaller than on non-Chinese websites, which offer many more pages over HTTPS.

As for the usage of client-side security policies on HTTPS websites, we only found 8 websites using HSTS policies, and 206 websites having secure cookies. Interestingly, we noticed that websites are more likely to enable other client-side security policies when they have HTTPS implemented. For example, 69.8% of Chinese HTTPS website also make use of HTTP-Only Cookies, a fraction much higher than the one presented earlier in Table 1 (43%).

As for the distribution of HTTPS websites, we didn't find any strong correlation between a website's popularity and its adoption of HTTPS (as shown in Figure 1). To better understand the Chinese websites using HTTPS, we categorized the 672 HTTPS websites, using McAfee's TrustedSource Web Database [10]. The number of HTTPS websites and the percentage of HTTPS websites having secure cookies over the top 10 categories are presented in Figure 2. One can see that, the financial websites are doing better than websites in all other categories, with 110 financial websites having HTTPS implemented, and about half of them using secure cookies.

With only 6.7% of websites having HTTPS implemented, the adoption rate of HTTPS on the Chinese web is much lower, compared to the global statistic of 27% of the Alexa top 1 million websites already used HTTPS in 2010 [24]. Although HTTPS is good for security, many websites are reluctant to migrate to HTTPS, due to several concerns, such as, SSL/TLS performance overhead, and operational costs. Beside these common concerns shared by websites globally, Chinese websites have to consider a China-specific issue when deciding whether to adopt HTTPS, which is the lack of HTTPS support from Baidu.

	% of Non-Chinese HTTPS websites	% of Chinese HTTPS websites	Example findings
Insecure SSL/TLS Implementation	70.9%	84.1%	SSL 2.0 enabled, insecure renegotiation (passport.baidu.com)
Post-to-HTTPS Forms	31.8%	21.4%	`http://wallet.tenpay.com/web` (tenpay.com)
Mixed-content Inclusion	20.7%	10.6%	`https://login.xiu.com/` (xiu.com)
HSTS Policy	3.6%	1.3%	`max-age=31536000` (alipay.com)
Secure Cookies	19.9%	30.7%	`CAMToken=h4Kei5MyrK3DzSNmB iVNr8skoJs=; Path=/; Secure` (hangseng.com.cn)

Table 2: Assessment overview for Chinese HTTPS websites

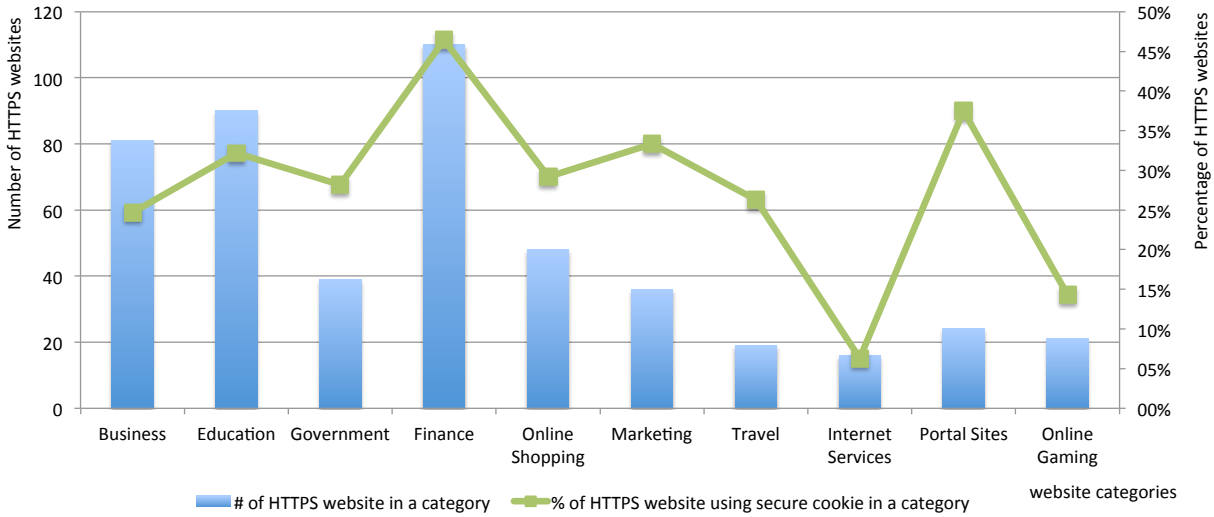

Figure 2: Distribution of Chinese HTTPS websites over top 10 categories

As explained in Baidu's official search engine optimization (SEO) guide [3], Baidu does not have a good support for HTTPS, thus Chinese websites are recommended to avoid the use of HTTPS. If HTTPS is essential for a website, the website should try to make the HTTPS webpages also available over HTTP, in order to get indexed by Baidu.

To further support our claim about Baidu's partial liability for the low HTTPS adoption on the Chinese web, we compare the search results from Baidu and Bing, for the 672 Chinese HTTPS websites. While we obtained about 2,000 HTTPS URLs for 296 (44%) HTTPS websites during data collection with Bing Search API, we didn't get any HTTPS URLs from Baidu when using the same queries like `site:example.com` to search for the popular webpages of each HTTPS website.

4.4 Usage of KNET trusted website certificate

One of the benefits of implementing HTTPS is website authenticity, which can help users to identify legitimate websites and prevent some phishing attacks. With only 605 Chinese websites having HTTPS implemented, and 13% of them using a self-signed certificate, it is challenging for Chinese Internet users to identify phishing websites by checking SSL certificates. In order to protect their users from phishing attacks, many Chinese websites opt for a Chinese-specific approach for website authentication by using trusted website certificate issued by KNET (owned by CNNIC) [7].

KNET certifies a website based on its ICP license, registration information in industrial and commercial bureau, and organization information. Certified websites are recorded in the "National Trusted Site Database Open Platform", which is used by various client-side applications to help users identify trusted websites. These applications include search engines (Sogou, SOSO and Bing), browsers (Sougou, Maxthon, Taobao, Ali, QQ and 114la), anti-virus software (Jinshan), and IE browser plugins. Considering the large amount of phishing websites on the Chinese web [18], and the various concerns about HTTPS implementations, we think that using a trusted website certificate is a practical and effective way for Chinese websites to defend themselves against phishing attacks.

By querying the trusted site database, we found that only 21.8% (2,182) of the investigated Chinese website are using KNET's trusted website certificate. Similar to the distribution of HTTPS websites, certified trusted websites mostly belong to categories such as government, business, finance, and online shopping.

5. IDENTITY LEAKAGE

Online data leakage incidents are usually caused by cyber attacks. For example, in December 2013, Chinese hackers leaked over 20 million hotel reservation records containing customer information [1]. Careless or improper man-

Category	#/% of websites	#/% of files	Example findings
Government	247 (41%)	1,592 (43.2%)	List of candidates at Guangzhou bureau of industry and commerce (广州市工商局体检名单) (gzaic.gov.cn)
Education	237 (39.3%)	1,543 (41.9%)	List of registrants in Zhejiang University (浙江大学报名点名单) (zju.edu.cn)
Others	119 (19.7%)	545 (14.9%)	List of participants in racewalking (竞走锦标赛人员名单) (sport.org.cn)

Table 3: Distribution of websites leaking spreadsheet files containing ID numbers

agement of sensitive data, however, may also lead to incidents. For instance, a Chinese university inadvertently made 2,000 students' identity numbers and bank accounts available, through files located on their website [2]. In this section, we investigate the issue of inadvertent data leakage. More specifically, we check whether a website is hosting spreadsheet files containing Chinese identity numbers (henceforth ID numbers), that can be obtained simply via search engines.

For each Chinese website in our dataset, we query both Google and Baidu with "身份证(literally means ID card) `site:example.cn filetype:xls`" in order to obtain the first hundred search results. Then, for each result, we examine the description part using regular expressions and the Chinese ID checksum algorithm [31], in order to find valid ID numbers. If the description part of a search result contains a valid ID number, then we claim that the search result (spreadsheet file) leaks private ID information. At the end of this process, we found 2,496 spreadsheet files hosted on 548 websites through Google searches, and 1,422 spreadsheet files hosted on 315 websites through Baidu searches.

In total, 3,680 spreadsheet files containing ID numbers were found, located on 603 (6%) Chinese websites. After categorizing these 603 websites, we found that the majority of these files are hosted on governmental and educational websites, as shown in Table 3. By examining the file names and descriptions from the results of the two search engines, we found that these spreadsheet files containing ID numbers are typically used for announcing the results of various activities, such as recruitment in government, enrollment in universities, and awarding prizes and scholarships.

The issue of inadvertent data leakage when publishing newsletters, is mainly due to the lack of privacy awareness in Chinese organizations and the lack of legal protection for private information in China. Although such inadvertent mistakes are not malicious in nature, the leaked IDs can be used by criminals for various illegal activities, such as selling them to teenagers for online gaming registration (to evade the anti-addiction system), using them for train ticket scalping (a valid ID is required for buying train tickets online), and operating malicious online shops with impersonated identities.

Moreover, some leaked spreadsheet files also contain other types of personal information such as phone numbers, occupations and addresses. A motivated adversary can leverage all these private information to conduct social engineering attacks against related individuals, and subsequently to launch targeted attacks against associated organizations. For example, some leaked spreadsheet files contain recruitment information for government, which could, in principal, be used by advanced persistent threat actors [22] to attack the Chinese government.

We consider this is a severe privacy issue on the Chinese web. If an ID number or phone number must be made available for identification or authentication, organizations can straightforwardly mask some digits in the number (e.g., 123456******1234 for an ID number, 138****1234 for a phone number), in order to protect private information. Considering that the last digit of an ID number is a checksum, it is better to also mask that digit (e.g., 12345619****567*) instead of masking only 4 birthday digits (as used for printing an ID number on a train ticket), in order to obtain higher anonymity and prevent brute-forcing the ID number based on the ID checksum algorithm.

To quantify websites that are using masked digits for protection, we also searched for masked ID numbers when examining results from search engines. In total, we found 83 websites using masked digits when publishing spreadsheet files. Not surprisingly, 9 of them are financial websites, including the big four Chinese banks. This shows that financial organizations put more effort in protecting themselves and their users.

6. RELATED WORK

As the web becomes more complex and popular, security and correctness become ever more crucial attributes of web applications. A variety of methods and techniques have been proposed to test web applications [21], and most of them are designed to detect specific vulnerabilities and errors such as SQL Injections and XSS attacks.

To the best of our knowledge, this paper is the first that attempts to analyze the security of the Chinese web from the aspect of the adoption of defense mechanisms. Zhuge et al. investigated China's online black markets through structure modeling and empirical analysis, showing that the size of the Chinese underground economy is astonishing [20]. Alarifi et al. presented a large-scale evaluation of popular Arabic websites, by analyzing malicious webpages from 7,000 domains using web scanner APIs like Google Safe Browsing [13]. Another regional web security assessment is presented in [27], where the authors analyzed the malicious servers in the `.nz` domain.

7. CONCLUSION

While the web gains more popularity in China, it is important for Chinese websites to identify their weaknesses and adequately protect themselves. To this end, in this paper, we present a security analysis of the top 10,000 Chinese websites. By investigating the usage of client-side security policies, and assessing the HTTPS implementations on the Chinese web, we observed that the majority of websites lack

support for client-side security policies, and that the statistics of vulnerable HTTPS implementations of Chinese websites are also worrisome. Moreover, we found that 6% of websites are leaking Chinese ID numbers through spreadsheet files that can be obtained by simple searches in Google and Baidu. We hope that our study can raise the security and privacy awareness among Chinese Internet users, and help Chinese websites to adopt client-side security policies and implement HTTPS in a secure way.

Acknowledgements

We want to thank the anonymous reviewers for their valuable comments. This research was performed with the financial support of the Prevention of and Fight against Crime Programme of the European Union (B-CCENTRE), the Research Fund KU Leuven, iMinds, IWT, and by the EU FP7 projects WebSand, NESSoS and STREWS.

8. REFERENCES

[1] The 20 million hotel reversation records (in chinese). `http://net.chinabyte.com/2/12850502.shtml`.

[2] 2000 students' id number and bank acount leaked (in chinese). `http://tech.sina.com.cn/i/2012-03-23/07326867582.shtml`.

[3] Baidu SEO Guide V2.0 (in Chinese). `http://baiduseoguide.com`.

[4] Bing Search API. `http://datamarket.azure.com/dataset/bing/search`.

[5] CSDN data breach (in Chinese). `http://baike.baidu.com/view/7167245.htm`.

[6] HtmlUnit. `http://htmlunit.sourceforge.net/`.

[7] KNET Trusted Website Database (in Chinese). `http://t.knet.cn/`.

[8] SSL Pulse. `https://www.trustworthyinternet.org/ssl-pulse/`.

[9] sslyze. `https://github.com/iSECPartners/sslyze`.

[10] TrustedSource Web Database. `https://www.trustedsource.org/`.

[11] Understanding the TLS Renegotiation Attack. `http://www.educatedguesswork.org/2009/11/understanding_the_tls_renegoti.html`.

[12] Wooyun (in chinese). `http://www.wooyun.org/bugs/`.

[13] Abdulrahman Alarifi and AbdulMalik AI-Salman. Security analysis of top visited arabic web sites. In *15th International Conference on Advanced Communication Technology*. IEEE, 2013.

[14] Adam Barth. HTTP State Management Mechanism. *IETF RFC 6265*, 2011.

[15] Ping Chen, Nick Nikiforakis, Christophe Huygens, and Lieven Desmet. A Dangerous Mix: Large-scale analysis of mixed-content websites. In *Proceedings of the 16th Information Security Conference*, Dallas, USA, 2013.

[16] Jeremy Clark and Paul C. van Oorschot. SoK: SSL and HTTPS: Revisiting Past Challenges and Evaluating Certificate Trust Model Enhancements. In *IEEE Symposium on Security and Privacy*, pages 511–525, 2013.

[17] CNNIC. The 32th China Internet Development Statistic (in Chinese).

[18] CNNIC and APAC. Global Chinese Phishing Sites Report. `http://www.cnnic.cn/gywm/xwzx/rdxw/rdxx/201305/W020130531616450986485.pdf`, 2013.

[19] Microsoft IEBlog. Declaring Security. `http://blogs.msdn.com/b/ie/archive/2009/06/25/declaring-security.aspx`, 2009.

[20] Zhuge Jianwei, Gu Liang, and Duan Haixin. Investigating China's Online Underground Economy. *Conference on the Political Economy of Information Security in China*, 2012.

[21] Yuan-Fang Li, Paramjit K. Das, and David L. Dowe. Two decades of Web application testing - A survey of recent advances. *Information Systems*, 43(0):20 – 54, 2014.

[22] Mandiant. The Advanced Persistent Threat. 2010.

[23] Moxie Marlinspike. New Tricks for Defeating SSL in Practice. 2009.

[24] Ivan Ristić. Internet SSL Survey 2010. In *Black Hat USA 2010*, 2010.

[25] Juliano Rizzo and Thai Duong. CRIME: Compression Ratio Info-leak Made Easy. In *ekoparty Security Conference*, 2012.

[26] D. Ross and T. Gondrom. HTTP Header Field X-Frame-Options. *IETF RFC 7034*, 2013.

[27] Christian Seifert, Vipul Delwadia, Peter Komisarczuk, David Stirling, and Ian Welch. Measurement Study on Malicious Web Servers in the .nz Domain. In *Information Security and Privacy*, volume 5594. Springer, 2009.

[28] Sid Stamm, Brandon Sterne, and Gervase Markham. Reining in the web with content security policy. In *Proceedings of the 19th international conference on World wide web*, WWW '10. ACM, 2010.

[29] Brandon Sterne and Adam Barth. Content Security Policy 1.0. *W3C Candidate Recommendation*, 2012.

[30] Tom van Goethem, Ping Chen, Nick Nikiforakis, Lieven Desmet, and Wouter Joosen. Large-scale Security Analysis of the Web: Challenges and Findings. In *Proceedings of the 7th International Conference on Trust and Trustworthy Computing*, 2014.

[31] Wikipedia. Resident Identity Card. `http://en.wikipedia.org/wiki/Resident_Identity_Card_(PRC)`.