Accessing Google Scholar under Extreme Internet Censorship: A Legal Avenue

Zhen Lu
Tsinghua University, VMware
Beijing, China

Zhenhua Li, Jian Yang
Tsinghua University
Beijing, China

Tianyin Xu
UCSD
CA, USA

Ennan Zhai
Yale University
CT, USA

Yao Liu
SUNY Binghamton
NY, USA

Christo Wilson
Northeastern University
NY, USA

ABSTRACT

Internet censorship is pervasive across the world. However, in some countries like China, even *legal*, nonpolitical services (*e.g.,* Google Scholar) are *incidentally* blocked by extreme censorship machinery. Therefore, properly accessing legal Internet services under extreme censorship becomes a critical problem. In this paper, we conduct a case study on how scholars from a major university of China access Google Scholar through a variety of middleware. We characterize the common solutions (including VPN, Tor, and Shadowsocks) by measuring and analyzing their performance, overhead, and robustness to censorship. Guided by the study, we deploy a novel solution (called ScholarCloud) to help Chinese scholars access Google Scholar with high performance, ease of use, and low overhead. This work provides an insider's view of China's Internet censorship and offers a legal avenue for coexistence with censorship.

CCS CONCEPTS

• **Networks → Network performance evaluation**; **Network measurement**; *Middle boxes / network appliances*;

ACM Reference format:
Zhen Lu, Zhenhua Li, Jian Yang, Tianyin Xu, Ennan Zhai, Yao Liu, and Christo Wilson. 2017. Accessing Google Scholar under Extreme Internet Censorship: A Legal Avenue. In *Proceedings of Middleware '17, Las Vegas, NV, USA, December 11–15, 2017,* 7 pages.
DOI: 10.1145/3154448.3154450

1 INTRODUCTION

Internet censorship, despite being highly controversial, is pervasive across the world mainly for political reasons. Some countries such as China, Iran, Russia, and Thailand wield extreme Internet censorship that even block *legal*, nonpolitical Internet services [1, 4, 20, 44], *e.g.,* Google Scholar. In these countries, easily accessing legal Internet services under extreme censorship becomes a critical problem.

As one of the largest research and education communities in the world, China's scholars unfortunately cannot access Google Scholar—the world's largest academic search engine. This creates enormous obstacles to their academic activities. Google Scholar is blocked by China's national-scale censorship firewall, known as the Great Firewall (GFW), as it is under the *google.com* domain.[1] Over the past decade, the GFW has become the most extensive and advanced Internet surveillance and control system in the world, adopting methods including IP blocking, DNS poisoning, URL filtering, packet filtering based on deep packet identification (DPI) and active probing [1–3, 8, 12, 17, 18, 26, 36, 41, 45, 48, 49, 51, 53, 54].

Unfortunately, Google Scholar is *incidentally* blocked in China: in fact, Google Scholar is principally considered a legal service by the relevant governmental regulators of China [30, 43, 46] but suffers from collateral damage of the GFW. The GFW is only responsible for technical blocking in China's complex Internet censorship system; non-technical policy and regulation are operated by government agencies, mainly including MIIT, TCA, MPS, and MSS (§2 explains the organizational structures of China's Internet censorship ecosystem in detail). The contradictory blocking of Google Scholar (and other ostensibly legal online services) stems from a critical fact that has been neglected in literature: *the behavior of the GFW is not always consistent with the policies of the Chinese government (and both evolve over time).*

To understand the above issue, we conduct a case study on how scholars (including 371 faculty members and students) at Tsinghua University, one of the top academic institutions in China, bypass the GFW to access Google Scholar through a variety of middleware. Interestingly, despite Google Scholar being blocked by the GFW, 26% of scholars in our survey are able to regularly bypass the GFW to access Google Scholar. We study the common bypass solutions: VPN (including both native VPN and OpenVPN), Tor, and Shadowsocks, by measuring their performance, overhead, and robustness against censorship. Our major findings are summarized as follows:

- Native VPN based on layer-2 tunneling protocols (PPTP and L2TP) is supported by most OSes. Our measurements show that it is currently robust to the GFW with a low packet loss rate (0.21% in average).[2] Unfortunately, since it forwards all traffic to remote VPN servers outside China, it significantly increases

Middleware '17, Las Vegas, NV, USA
© 2017 ACM. 978-1-4503-5200-0/17/12…$15.00
DOI: 10.1145/3154448.3154450

[1] Internet services under the Google domain have been blocked in China since 2010, including Search, Gmail, Maps, Docs, Drive, Google+, Sites, and Picasa [14].

[2] The GFW's censorship towards VPNs has changed significantly over the years. During 2012–2015, VPNs were extensively blocked by the GFW [33, 34]. Starting from 2015, registered VPN services have become legal and pervasive in China [5, 13]. In contrast, unregistered VPN services are considered illegal and many of them have been shut down by the government agencies [42].

access latency to domestic Internet services. Hence, users have to frequently and manually reconfigure their network connections.

- OpenVPN implements both layer-2 and layer-3 VPN tunnels with multiple customizations for security and resilience (*e.g.,* traffic redirection). Its performance is similar to that of native VPN when accessing Google Scholar. However, OpenVPN requires users to install extra client software and go through complicated configurations, considerably decreasing its usability.
- Tor requires the installation of a Tor client which directs browser traffic through a free, volunteer network of relays and bridges. Consequently, we find that it suffers from large client-side overhead and the longest first-time page load time (PLT) between 13 and 20 sec (= seconds). Still worse, even when using the latest *meek* obfuscation protocol, Tor is severely censored by the GFW, indicated by the high packet loss rate (4.4% in average).
- Shadowsocks [9] sets up an encrypted connection between a local proxy client and a remote proxy server on top of the layer-4 TLS protocol. The project was banned in China in 2015, but the software has been widely spread since 2012. We find Shadowsocks suffers from the longest average PLT (3.7 sec) because it imposes an extra TCP connection for user/password authentication into each HTTP session. We also find that it is vulnerable to the GFW [6, 7], with an average packet loss rate of 0.77%.

Our study shows that none of the common solutions have provided satisfactory user experiences for accessing Google Scholar. Note that most scholar users do not have a background of computer science or engineering; it is unrealistic to expect these users to manage advanced software setups and configurations. To effectively help scholars in China access Google Scholar, we deploy a novel solution, ScholarCloud, with high performance, ease of use, and low overhead. This is achieved by the following endeavors:

- ScholarCloud uses a **split-proxy architecture** that encapsulates complicated proxy configuration and software installation (required by OpenVPN and Shadowsocks) into an easy-to-use domestic proxy.
- We implement **message blinding** that obfuscates the encrypted messages by encoding them into formats that are not recognized by the GFW. This achieves a low packet loss rate (0.22% in average) and short PLT (1.3 sec in average).
- We focus on helping users access legal services that are incidentally blocked by the GFW. Therefore, we have **legalized ScholarCloud** through registration at government agencies with a *visible* whilelist of services.

We describe the ScholarCloud system in more detail in §3. It was launched in Jan. 2016 and has served more than 2000 users, with 700 users online every day at present. The service can be accessed via *http://scholar.thucloud.com.* The system is hosted on top of two regular VM (virtual machine) servers; its daily operational cost is merely 2.2 USD. This work provides an insider's view of China's Internet censorship (§2), which is mostly missed or even misunderstood in prior work. Most importantly, ScholarCloud offers a legal avenue for accessing Internet content under extreme censorship. Despite its limitations (§6), it is currently the most practical and sustainable solution built inside the wall.

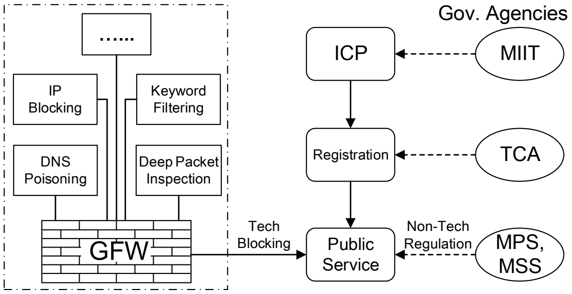


Figure 1: The bilateral ecosystem of China's Internet censorship: the GFW for technical blocking and the relevant government agencies for non-technical regulation.

2 INTERNET CENSORSHIP IN CHINA

China's Internet censorship ecosystem is bilateral—it consists of the GFW for technical blocking and the relevant government agencies for non-technical regulation. To the best of our knowledge, these two components do not operate synchronously; inconsistencies are often observed (*e.g.,* whether or not to prohibit access to Google Scholar). Figure 1 depicts the relationship between the two components. Most of the prior work studies the GFW in terms of its censorship techniques; however, little correctly understand the government agencies and their regulation polices. This section focuses on their organizational structure and respective roles.

At present, there are mainly four government agencies collaboratively regulating public Internet services in China: the Ministry of Industry and Information Technology (MIIT), the Ministry of Public Security (MPS), the Ministry of State Security (MSS), and the Telecommunication Administration (TCA) agencies located in each city [11, 37, 38]. Specifically, any organization or company who attempts to deliver digital content across the Internet in China is termed an Internet Content Provider (ICP), such as Alibaba, Baidu, and Tencent. Any ICP who intends to provide public services is required to *register* the service at the TCA agencies in the corresponding city. MIIT is responsible for formulating legislation, guiding the TCA agencies, and maintaining a centralized database of registered ICPs, hosted at *http://www.miitbeian.gov.cn.*

Registration is a manual process. It involves recording and subsequently verifying the ICP's service name, service type, domain name, responsible person, and other critical information (which typically takes weeks to months) [38–40]. If a registered ICP is considered illegal (providing illegal Internet services), the MPS and MSS are responsible for quickly shutting down the service by blocking the domain name and even arresting the responsible person.

Different from the GFW's aggressive technical blocking, MPS and MSS adopt a relatively *conservative* approach to regulating Internet services (*i.e.,* they usually block an Internet service after extensive investigation and evidence collection). This is partially because non-technical regulations are difficult to automate and thus take time to apply in practice. For example, many ICPs that provide VPN services do not register at the TCA agencies. Although *in theory* the MPS and MSS have the right to shut down any Internet services provided by an unregistered ICP, detecting, identifying, and localizing them is not an easy job given the open and distributed nature of the

Internet. In reality, a number of transnational corporations have their branches in China and they often provide unregistered VPN services to their employees [35]. In this case, if the MPS or MSS were to simply block all VPN services, this would create disputes and hinder economic development [1].

In this work, we demonstrate using ScholarCloud that it is possible and feasible to provide high-performance, usable services to access legal Internet content, even under extreme censorship. On one hand, the service needs to bypass the technical blocking of the GFW; on the other hand, the service has to adhere to the policies and regulations enforced by government agencies.

3 THE SCHOLARCLOUD SYSTEM

ScholarCloud provides a platform to help Chinese scholars access legal Internet services incidentally blocked by the GFW. It is designed for scholar users who may not have a technical background, and does not require any software installation. ScholarCloud works transparently to its users—they only need to configure their browsers using a proxy auto-config (PAC [32]) file generated by our system. All of the underlying connection tunneling is then automatically handled and the users do not even notice the existence of our system. The PAC file only diverts traffic for a small whitelist of legal (but incidentally blocked) domains to our proxy; all other traffic is routed to the Internet normally. Figure 2 illustrates the architectural difference between ScholarCloud and four other solutions. ScholarCloud achieves high performance and usability, as well as low overhead by the following main endeavors.

Split-proxy architecture and configuration automation. Our ScholarCloud system adopts a split-proxy architecture—a domestic proxy located inside China and a remote proxy outside China. This design is inspired by Shadowsocks which also has a dual-proxy architecture. However, Shadowsocks places a local proxy on every client device (which requires software installation and extensive configuration). In contrast, ScholarCloud encapsulates the per-client proxies into a logically centralized, domestic proxy server that automates the vast majority of configuration. Users simply need to change one setting in their browser to use a PAC file supplied by ScholarCloud to begin using the service.

Message blinding. ScholarCloud aims to bypass the GFW in a robust and efficient way. Prior work has shown that the GFW is able to identify and block traffic using well-known encryption and obfuscation schemes (*e.g.,* HTTPS, AES, and *meek*) [6, 45]. To address this issue, ScholarCloud blinds the messages between the domestic proxy and the remote proxy by adopting a custom encoding approach which is confidential and not recognized by the GFW. There are a number of encoding approaches that can potentially meet our requirements [16, 27]. In our experiments, we notice that even using a simple but non-public algorithm like byte mapping ($f : [0, 2^8) \rightarrow [0, 2^8)$) can successfully blind the encrypted messages from the GFW's inspection, thus bypassing the GFW. Therefore, by making the encrypted messages blind to the GFW, we give our system high resistance to the Internet censorship.

Since we control both the domestic and remote proxy, we can change our blinding mechanism at any time without impacting users. This gives ScholarCloud the ability to adapt to GFW's reactions to our blinding scheme, *i.e.,* we are agile against future changes in the GFW. In contrast, Tor cannot easily do this because it would require many independent relays to upgrade; similarly, Shadowsocks can only adapt if users upgrade their client software.

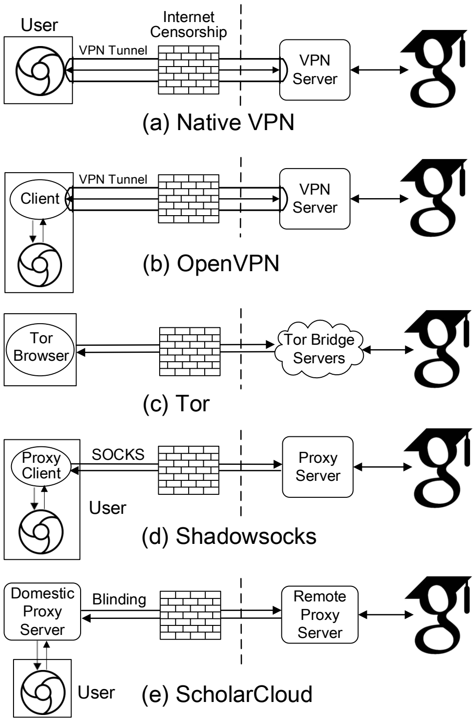


Figure 2: Architectural overview of the studied solutions.

Data security and privacy. During the three phases in Figure 2(e) (*i.e.,* User ↔ Domestic Proxy, Domestic Proxy ↔ Remote Proxy, and Remote Proxy ↔ Target Web Server), if a message is already encrypted with HTTPS (by the user's browser and the target web server), ScholarCloud will not encrypt it again; otherwise, ScholarCloud creates an encrypted tunnel between the domestic and remote proxy using HTTPS. Note that there is a potential privacy risk when using ScholarCloud, as it can observe the content of unencrypted packets. However, this risk also exists with VPN services, Tor exit nodes, and Shadowsocks proxies; ScholarCloud does not increase the privacy risks to users.

Service legalization. We focus on helping users access legal services that are incidentally blocked by the GFW. Therefore, we have legalized ScholarCloud by registering it with the relevant government agencies (with China's ICP Reg. #15063437). Specifically, besides providing a series of information about our service (as mentioned in §2), we start a company as the business entity of our ScholarCloud service and register at the responsible TCA agency. The registration requires the following documents: (1) biometric document of the legal representative of the company; (2) documentation of the ScholarCloud service with text, screenshots, and real-world usage videos, as well as a workable user guide.

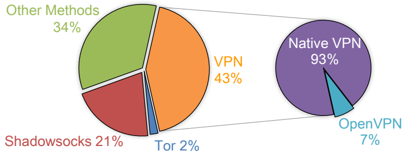


Figure 3: Distribution of methods for accessing Google Scholar adopted by scholars in Tsinghua University.

In addition, ScholarCloud only redirects those user traffic corresponding to a *visible* whitelist of legal services (that are incidentally blocked). Only when a service (domain) appears in the whitelist, ScholarCloud redirects the traffic to bypass the GFW; otherwise, it does not modify the traffic at all. Therefore, government agencies can examine which legal services are defined by us, and request that we alter the whitelist on demand.

4 MEASUREMENT STUDY

In this section, we first introduce our user survey results to understand the common practices of accessing Google Scholar in China. Next, we present our methodology for comparing different GFW circumvention techniques. Finally, we report the measurement results as well as our analysis and insights.

4.1 Common Practices

To understand how Chinese scholars access Google Scholar, we posted a survey through the bulletin board system (BBS) of Tsinghua University in July 2015, and received 371 results from faculty members and students. Most of the participants were not from the departments of computer science or engineering. Note that ScholarCloud had not been deployed in 2015, and thus was not an option at the time of the survey. Figure 3 shows the distribution of methods for accessing Google Scholar reported by the participants. 26% of the scholars reported bypassing the GFW and accessing Google Scholar. Among them, 43% reported utilizing VPNs (93% using native VPN and 7% using OpenVPN), 2% used Tor, 21% used Shadowsocks, and the remaining 34% adopted other solutions (*e.g.,* other web proxies such as Free Gate [10], or modifying their system's hosts file). Therefore, our measurements focus on the four most common solutions, as well as ScholarCloud (c.f., Figure 2).

4.2 Methodology

To comparatively study the performance of the five access methods in a fair manner, we conduct controlled benchmark experiments during Feb.–Apr., 2017. In all the experiments, the client is a common laptop (ThinkPad T440s) located at Tsinghua University (Beijing, China) inside CERNET (China Education and Research Network). The client uses 64-bit Windows 8.1. Except for Tor, that requires the dedicated Tor browser (version 6.5), we use the Chrome web browser (version 56.0, 64-bit) to access Google Scholar. In all cases, we automate the web browser to send HTTP requests for the home page of Google Scholar every 60 sec, (hence two consecutive accesses do not overlap) and each experiment lasts for a whole day.

On the server side, except for Tor where the remote servers are assigned by the Tor network, we employ a VM server rented from Aliyun ECS located at San Mateo, CA, USA. The server has a single-core Intel Xeon CPU (2.3 GHz), 1 GB DDR3 memory, 50 GB HDD disk storage, and an auto-scale Internet connection (with maximum bandwidth of 100 Mbps). The server OS is 64-bit Ubuntu 14.04. For ScholarCloud, we rent another VM to function as the domestic proxy located in Tsinghua University with the same configuration.

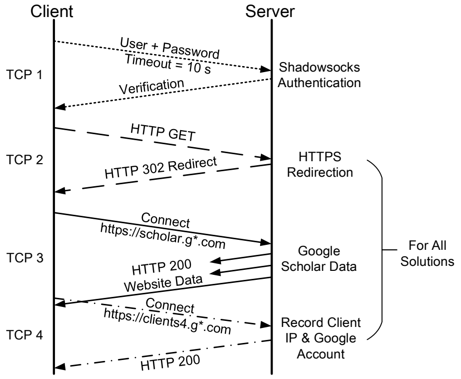


Figure 4: Detailed client-server interactions during the whole HTTP session to access Google Scholar of the studied methods. Note that the four TCP connections (TCP 1–4) are *not* present in all cases: TCP-1 (for user/password authentication) is only set up by Shadowsocks; TCP-2 (for HTTPS redirection) only appears when the client sends an HTTP request rather than an HTTPS request; TCP-4 (for recording the client's IP and Google account) only appears when the client accesses Google Scholar for the first time. TCP-3 (for real Google Scholar data exchange) appears in all cases.

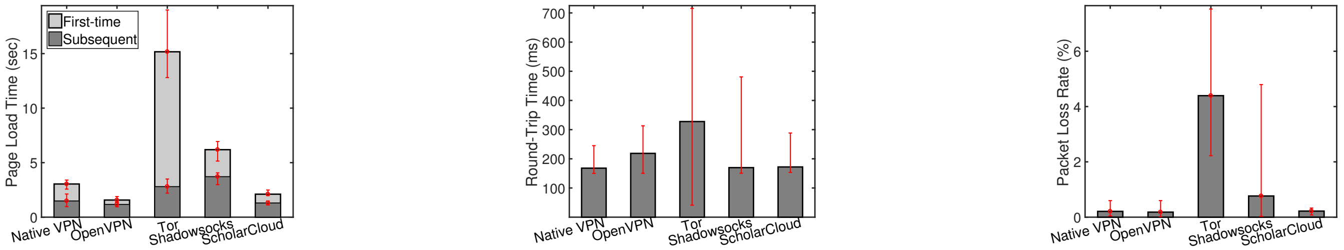
In our experiments, different access methods exploit different communication and encryption protocols. For native VPN, we use the PPTP protocol based on pptpd [3]. For OpenVPN, we adopt the layer-3 implementation, and use the Easy-RSA tool to create the PKI certificates and keys. For Tor, we employ the latest *meek* obfuscation protocol for connecting to bridge servers. For Shadowsocks, we encrypt the data connection between the proxy client and the proxy server based on the AES-256-CFB implementation. Finally, for ScholarCloud, we blind the messages using our custom scheme, so as to maximize its robustness to the censorship of the GFW.

4.3 Measurement Results

We report measurement results from the perspectives of performance, robustness to blocking, and overhead. Specifically, we examine the following metrics:

- *Page load time* (PLT) denoting user-perceived web experience;
- *Round trip time* (RTT) representing network-level efficiency;
- *Packet loss rate* (PLT) indicating robustness to censorship;
- *Client-side overhead*: network traffic, CPU, and memory usages.
- *Scalability* illustrating how many requests a solution can simultaneously support without obvious performance degradation.

[3] We also tested with other VPN protocols (L2TP and IPSec) based on implementations of xl2tpd, openswan, and ppp, and observed similar performance to PPTP.



(a) Page load time.

(b) Round trip time.

(c) Packet loss rate.

Figure 5: Performance and robustness of different access methods. The error bars show the max and min values observed.

To facilitate the analysis of measurement results, we plot in Figure 4 the interactions between the client and the server in the whole HTTP session for an access to Google Scholar.

Page load time (PLT). As shown in Figure 5a, PLTs can be dramatically different when the page is loaded for the *first time* and *subsequent* accesses. The former is much longer than the latter for three main reasons. First, there is no local DNS cache when accessing a web page for the first time, so the DNS resolution procedure must be performed. Second, there is no local content cache for the web page, so the web client has to setup required data connections to all the involved web servers to fetch corresponding content. Third, the Google Scholar server needs to record the client's IP address and Google account, as demonstrated by the fourth TCP connection (TCP 4) in Figure 4.

Tor has the largest initial PLT, since its connection setup process involves interactions with multiple bridges and relays. The first-time PLT (15 sec) is 5.4 times longer than the normal PLT (2.8 sec) on average. In the worst case, the first-time PLT is close to 20 sec.

Native VPN and OpenVPN have similar normal PLTs between 1.2 and 1.5 sec, while Shadowsocks has much longer PLT (3.7 sec on average). In order to understand the root cause, we delve into the implementation of Shadowsocks at the source-code level [9]. We find that Shadowsocks imposes an extra TCP connection for user/password authentication in the beginning of each HTTP session, as demonstrated by TCP 1 in Figure 4. Additionally, the default configuration of keep-alive timeout in Shadowsocks is 10 sec, *i.e.,* Shadowsocks reinitializes the authentication procedure if there is no request passing through the connection in 10 sec. This makes the communication more secure but considerably prolongs PLT.

Although ScholarCloud and Shadowsocks adopt a similar dual-proxy architecture, the former's PLTs (first-time 2.1 sec and subsequent 1.3 sec on average) are remarkably shorter than the latter's for two main reasons. First, ScholarCloud does not impose an extra TCP connection for user/password authentication in each HTTP session. Second, ScholarCloud's message blinding mechanism greatly reduces the packet loss rate (see Figure 5c) which also influences the PLT. As a result, ScholarCloud achieves short PLTs.

Round trip time (RTT). It is known that RTT is correlated to PLT. In this section, we quantify the impact of RTT on PLT. According to Figure 5b, RTT has stronger correlations with the first-time PLT than the normal PLT. For example, Tor bears the longest first-time PLT (15 sec in average) as well as the longest RTT (330 ms in average). This can be explained by the fact that the first-time access to a web page requires more round trips than subsequent accesses, thus making RTT have larger influence on the first-time PLT.

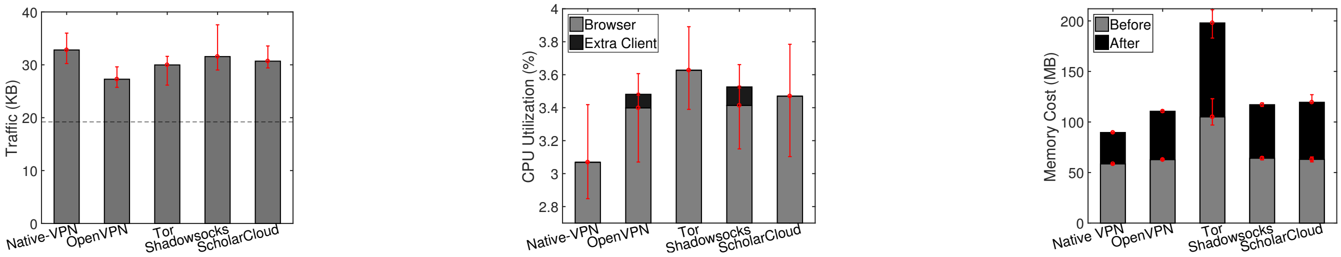
Packet loss rate (PLR). As explained in §1, the GFW exploits a variety of techniques for Internet censorship, including IP blocking, DNS poisoning, URL filtering, packet filtering, and so forth. Many of these techniques result in packet losses over the end-to-end connection. Therefore, we use PLR as a key indicator of robustness against the censorship of the GFW.

As illustrated in Figure 5c, Tor is severely impacted by the censorship of the GFW, with the highest PLR (4.4%) on average. Shadowsocks is also vulnerable to the censorship of the GFW with the average PLR being 0.77%. In comparison, when using Tor or Shadowsocks in the US to access Google Scholar, we observe that the PLR usually stays below 0.1%. Over the past a few years, a number of techniques specific to Tor and Shadowsocks have been added into the GFW, such as deep packet inspection and active probing [17, 31, 49]. Therefore, we do not use these methods as the building blocks of ScholarCloud.

Both native VPN and OpenVPN are robust to the censorship of the GFW, with low PLR being around 0.2% on average, similar to the PLR of accessing non-blocked US websites (*e.g.,* Amazon) from a web client in China. We do not build ScholarCloud upon VPN because the government policies towards VPN keep changing in China [35, 42]. On the other hand, ScholarCloud's message blinding achieves a similar PLR (0.22% on average) compared with VPN.

Client-side overhead. Figure 6 illustrates the client-side overhead of different access methods in terms of network traffic, CPU, and memory, respectively. Compared with accessing non-blocked web sites, bypassing GFW requires extra network traffic for tunneling, encrypting, or obfuscating the exchanged data. As indicated in Figure 6a, during a direct access to Google Scholar from a web browser in the US, the network traffic amounts to 19 KB on average (the dotted line). When bypassing the GFW to access Google Scholar, OpenVPN adds the least traffic overhead (8 KB) while native VPN addes the most (14 KB). In general, none of the solutions, including ScholarCloud, exhibit significant traffic overhead.

The CPU utilizations of the web browser and the extra client software (if there is any) are listed in Figure 6b. Native VPN increases CPU utilization the least (3.07%) while Tor increases it the



(a) Network traffic.

(b) CPU usage.

(c) Memory usage.

Figure 6: Client-side overhead of different access methods. The error bars show the max and min values observed.

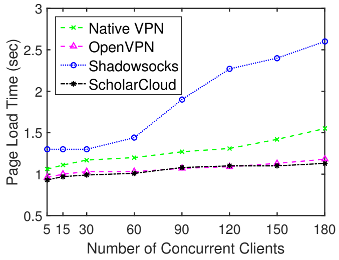


Figure 7: Scalability of each access method in terms of PLT.

most (3.62%). However, the increase percentage (18%) from 3.07% to 3.62% is not remarkable. Also, the extra CPU cost brought by the extra client software of OpenVPN and Shadowsocks is trivial.

For memory usage, as shown in Figure 6c, when the web browser is running but not being really used to access those blocked web content ("Before"), the Tor browser consumes nearly 70% more memory than Chrome. Also, when the web browser is actually used to access Google Scholar ("After"), native VPN consumes the least extra memory usage (30 MB) while Tor consumes the most (90 MB). The two observations consistently shows that the complication of Tor leads to obviously larger memory consumption.

Scalability. Figure 7 shows the scalability of native VPN, OpenVPN, Shadowsocks, and ScholarCloud. We do not measure the scalability of Tor because we are unable to control the Tor bridge servers. Sepcifically, we use the average page load time (PLT) as the major indicator of scalability as the number of concurrent clients increases, because PLT is the most important user-perceived web experience. We observer that Shadowsocks bears the worst scalability for its PLT sharply grows when the number of concurrent clients exceeds 60. In contrast, the PLTs of native VPN, OpenVPN, and ScholarCloud exhibit linear growth when there are more concurrent clients. Among the three linear cases, OpenVPN and ScholarCloud possess the best scalability with their PLTs growing gently.

5 RELATED WORK

After launching since 2000, the GFW has raised significant concerns. There has been extensive work that attempts to understand the blocking techniques adopted by the GFW and circumvent its censorship. As revealed by previous work, the GFW has deployed at least four types of techniques to restrict the activities of Internet users in China [53]: IP blocking [8, 54], DNS poisoning/hijacking [2, 3, 26], keyword filtering [12, 41], and deep packet inspection [6, 7, 45]. It is reported that 99% of the blocking behavior of the GFW occur at the border routers between China and the US [2, 12], which motivates the split-proxy architecture of ScholarCloud.

The deployment of Internet censorship systems has led to an everlasting arms race between blocking techniques and circumvention approaches [15, 16, 19, 21–24, 27, 28, 47, 52]. Many circumvention approaches that were once effective at bypassing the GFW are currently impaired by the GFW's new blocking techniques. One such example is Tor, which has become less effective after the GFW's deployed Deep Packet Inspection [17, 18, 25, 29, 31, 48–50]. In our survey (§4.1), Tor is only used by 2% of surveyed users.

Our work is complementary to the aforementioned work. We focus on helping users to access legal Internet services rather than complete censorship circumvention. In China, many valuable and beneficial Internet resources (*e.g.,* Google Scholar) are incidentally blocked. However, there has been little work on accessing these legal Internet services along with censorship.

6 LIMITATION AND DISCUSSION

Despite its real-world deployment and sound performance, ScholarCloud has its limitations. Its user group is mainly composed of China's scholars, mostly faculties and students in certain universities. Also, it is used for whitelisted websites rather than all blocked websites by the GFW. Hence, its proxy servers receive less traffic and interact with fewer content providers compared with VPN, Tor, and Shadowsocks. In addition, ScholarCloud is a web-based proxy solution so it cannot help the users access those non-HTTP(S) content. Essentially, the web-based design strategy is a double-edge sword—it greatly simplifies the configurations imposed on users, while inevitably restricts the application scenarios.

ScholarCloud demonstrates a practical and sustainable approach to help users access legal Internet services with the existence of extreme censorship, complementary to the solutions that rely on external services outside China. In addition, we hope that this paper can help the community better understand China's Internet censorship ecosystem, which provides opportunities for designing circumstance solutions to benefit a massive population of end users.

Acknowledgement. This work is supported by High-Tech R&D Program of China under grant 2015AA01A201, China NSF under grants 61471217, 61432002 and 61632020, US NSF under grant CNS-1618931, and CCF-Tencent Open Fund under grant RAGR20160105.

REFERENCES

[1] Daniel Anderson. 2012. Splinternet Behind the Great Firewall of China. *ACM Queue* 10, 11 (November 2012), 1–10.

[2] Anonymous. 2012. The Collateral Damage of Internet Censorship by DNS Injection. *ACM SIGCOMM Computer Communication Review (CCR)* 42, 3 (July 2012), 22–27.

[3] Anonymous. 2014. Towards a Comprehensive Picture of the Great Firewall's DNS Censorship. In *Proc. of USENIX FOCI*.

[4] Simurgh Aryan, Homa Aryan, and J. Alex Halderman. 2013. Internet Censorship in Iran: A First Look. In *Proc. of USENIX FOCI*.

[5] The Beijinger. 2015. China Ministry Says Foreign VPN Operators Must Register. (2015). https://www.thebeijinger.com/blog/2015/01/28/china-ministry-says-foreign-vpn-operators-must-register.

[6] Xiang Cai, Xin Cheng Zhang, Brijesh Joshi, and Rob Johnson. 2012. Touching from a Distance: Website Fingerprinting Attacks and Defenses. In *Proc. of ACM CCS*.

[7] Wei Chen, Chenyang Li, Jing Shen, Wei Zhang, and Geng Yang. 2016. Webpage fingerprint identification method aiming at specific website category. (2016). https://www.google.com/patents/CN105281973A?cl=en

[8] Richard Clayton, Steven J Murdoch, and Robert NM Watson. 2006. Ignoring the Great Firewall of China. In *Proc. of the 6th Workshop on Privacy Enhancing Technologies (PETS)*.

[9] Clowwindy. 2012. Shadowsocks project. (2012). https://github.com/shadowsocks.

[10] Global Internet Freedom Consortium. 2002. Free Gate web site. (2002). https://freegate.en.softonic.com/.

[11] The State Council. 1996. Interim Provisions of the People's Republic of China Governing International Interconnection of Computer-based Information Networks. http://www.cac.gov.cn/1996-02/02/c_126468621.htm. (1996). Decree No. 218 of the State Council.

[12] Jedidiah R. Crandall, Daniel Zinn, Michael Byrd, Earl Barr, and Rich East. 2007. ConceptDoppler: A Weather Tracker for Internet Censorship. In *Proc. of ACM CCS*.

[13] People's Daily. 2015. VPN Services Should Be Registered in China. (2015). http://legal.people.com.cn/n/2015/0128/c188502-26462092.html.

[14] David Drummond. 2010. A new approach to China: an update. https://googleblog.blogspot.com/2010/03/new-approach-to-china-update.html. (2010). The Official Google Blog. Retrieved March 24, 2010.

[15] Kevin P Dyer, Scott E Coull, Thomas Ristenpart, and Thomas Shrimpton. 2013. Protocol Misidentification Made Easy with Format-Transforming Encryption. In *Proc. of ACM CCS*.

[16] Kevin P Dyer, Scott E Coull, and Thomas Shrimpton. 2015. Marionette: A Programmable Network Traffic Obfuscation System. In *Proc. of the 24th USENIX Security Symposium*.

[17] Roya Ensafi, David Fifield, Philipp Winter, Nick Feamster, Nicholas Weaver, and Vern Paxson. 2015. Examining How the Great Firewall Discovers Hidden Circumvention Servers. In *Proc. of ACM IMC*.

[18] Roya Ensafi, Philipp Winter, Abdullah Mueen, and Jedidiah R. Crandall. 2015. Analyzing the Great Firewall of China Over Space and Time. In *Proc. of the 15th Privacy Enhancing Technologies Symposium (PETS)*.

[19] David Fifield, Chang Lan, Rod Hynes, Percy Wegmann, and Vern Paxson. 2015. Blocking-resistant communication through domain fronting. *Proc. on Privacy Enhancing Technologies (PETS)* (2015).

[20] Genevieve Gebhart, Anonymous Author, and Tadayoshi Kohno. 2017. Internet Censorship in Thailand: User Practices and Potential Threats. In *Proc. of the 2nd IEEE European Symposium on Security and Privacy (EuroS&P)*.

[21] Amir Houmansadr, Chad Brubaker, and Vitaly Shmatikov. 2013. The Parrot Is Dead: Observing Unobservable Network Communications. In *Proc. of S&P*.

[22] Amir Houmansadr, Giang TK Nguyen, Matthew Caesar, and Nikita Borisov. 2011. Cirripede: Circumvention Infrastructure using Router Redirection with Plausible Deniability. In *Proc. of ACM CCS*.

[23] Josh Karlin, Daniel Ellard, Alden W Jackson, Christine E Jones, Greg Lauer, David Mankins, and W Timothy Strayer. 2011. Decoy Routing: Toward Unblockable Internet Communication. In *Proc. of USENIX FOCI*.

[24] Sheharbano Khattak, Mobin Javed, Philip D Anderson, and Vern Paxson. 2013. Towards Illuminating a Censorship Monitor's Model to Facilitate Evasion. In *Proc. of USENIX FOCI*.

[25] Zhen Ling, Junzhou Luo, Wei Yu, Ming Yang, and Xinwen Fu. 2012. Extensive Analysis and Large-Scale Empirical Evaluation of Tor Bridge Discovery. In *Proc. of IEEE INFOCOM*.

[26] Graham Lowe, Patrick Winters, and Michael L. Marcus. 2007. *The Great DNS Wall of China*. Technical Report. New York University.

[27] Daniel Luchaup, Kevin P Dyer, Somesh Jha, Thomas Ristenpart, and Thomas Shrimpton. 2014. LibFTE: A Toolkit for Constructing Practical, Format-Abiding Encryption Schemes. In *Proc. of the 23th USENIX Security Symposium*.

[28] Xiapu Luo, Peng Zhou, Edmond WW Chan, Wenke Lee, Rocky K. C. Chang, and Roberto Perdisci. 2011. HTTPOS: Sealing Information Leaks with Browser-side Obfuscation of Encrypted Flows. In *Proc. of NDSS*.

[29] Akshaya Mani and Micah Sherr. 2017. HisTorε: Differentially Private and Robust Statistics Collection for Tor. In *Proc. of NDSS*.

[30] Mashable. 2017. Google Scholar might finally be Google's way back into China. (2017). http://mashable.com/2017/03/13/google-scholar-in-china/#uQIB4xtjFSqB.

[31] Jon McLachlan and Nicholas Hopper. 2009. On the risks of serving whenever you surf: vulnerabilities in Tor's blocking resistance design. In *Proc. of the 8th ACM Workshop on Privacy in the Electronic Society (WPES)*. 31–40.

[32] Netscape. 1996. Proxy auto-config (PAC) file. (1996). https://en.wikipedia.org/wiki/Proxy_auto-config.

[33] BBC News. 2015. BBC report: China blocks virtual private network. (2015). http://www.bbc.com/news/technology-30982198.

[34] China News. 2013. China News: It is much harder for Chinese netizens to get across Great Firewall since Nov 2012. (2013). http://news.creaders.net/china/2013/03/02/1238807.html.

[35] China News. 2017. MIIT: Cleansing Illegal VPN Services Will not Affect Normal Operation of Transnational Corporations. (2017). http://www.chinanews.com/cj/2017/01-24/8134818.shtml.

[36] Daiyuu Nobori and Yasushi Shinjo. 2014. VPN Gate: A Volunteer-Organized Public VPN Relay System with Blocking Resistance for Bypassing Government Censorship Firewalls. In *Proc. of NSDI*.

[37] State Council of China. 1997. Administrative Measures for Protection of the Security of International Internetworking of Computer Information Networks. (1997). http://www.mps.gov.cn/n2254314/n2254409/n2254443/n2254451/c4113546/content.html.

[38] State Council of China. 2000. Administrative Measures on Internet Information Services. (2000). http://www.gov.cn/gongbao/content/2000/content_60531.htm.

[39] The Ministry of Information Industry. 2005. Measures for the Administration of Record-Filing of Non-Profit-Making Internet Information Services. http://www.miit.gov.cn/n11293472/n11293877/n11301753/n11496139/11537734.html. (2005). Order No. 33 of the Ministry of Information Industry.

[40] The Ministry of Information Industry. 2010. Ministry of Industry and Information Technology: On the Further Implementation of the Site Record Information Authenticity Verification Program. http://www.miitbeian.gov.cn/state/outPortal/queryMutualityDownloadInfo.action?id=267. (2010). Order No. 64.

[41] Jong Chun Park and Jedidiah R. Crandall. 2010. Empirical Study of a National-Scale Distributed Intrusion Detection System: Backbone-Level Filtering of HTML Responses in China. In *Proc. of IEEE ICDCS*.

[42] South China Morning Post. 2017. China tightens Great Firewall by declaring unauthorised VPN services illegal. (2017). http://www.scmp.com/news/china/policies-politics/article/2064587/chinas-move-clean-vpns-and-strengthen-great-firewall.

[43] South China Morning Post. 2017. Google Scholar may be Google's first service to return to China, Chinese lawmaker reveals. (2017). http://www.scmp.com/news/china/policies-politics/article/2078173/google-another-step-closer-being-unblocked-china.

[44] John-Paul Verkamp and Minaxi Gupta. 2012. Inferring Mechanics of Web Censorship Around the World.. In *Proc. of USENIX FOCI*.

[45] Liang Wang, Kevin P Dyer, Aditya Akella, Thomas Ristenpart, and Thomas Shrimpton. 2015. Seeing through Network-Protocol Obfuscation. In *Proc. of ACM CCS*.

[46] Tech Web. 2017. Google Scholar is Likely to Return China. (2017). http://www.techweb.com.cn/internet/2017-03-14/2499526.shtml.

[47] Zachary Weinberg, Jeffrey Wang, Vinod Yegneswaran, Linda Briesemeister, Steven Cheung, Frank Wang, and Dan Boneh. 2012. StegoTorus: A Camouflage Proxy for the Tor Anonymity System. In *Proc. of ACM CCS*.

[48] Philipp Winter and Jedidiah R. Crandall. 2012. The Great Firewall of China: How It Blocks Tor and Why It Is Hard to Pinpoint. *;login:* 37, 6 (December 2012), 42–50.

[49] Philipp Winter and Stefan Lindskog. 2012. How the Great Firewall of China is Blocking Tor. In *Proc. of USENIX FOCI*.

[50] Philipp Winter, Tobias Pulls, and Juergen Fuss. 2013. ScrambleSuit: A Polymorphic Network Protocol to Circumvent Censorship. In *Proc. of the 12th ACM Workshop on Privacy in the Electronic Society (WPES)*.

[51] Joss Wright. 2012. *Regional Variation in Chinese Internet Filtering*. Technical Report. Oxford Internet Institute, University of Oxford.

[52] Eric Wustrow, Scott Wolchok, Ian Goldberg, and J Alex Halderman. 2011. Telex: Anticensorship in the Network Infrastructure. In *Proc. of USENIX Security*.

[53] Xueyang Xu, Z. Morley Mao, and J. Alex Halderman. 2011. Internet Censorship in China: Where Does the Filtering Occur?. In *Proc. of the 12th International Conference on Passive and Active Measurement (PAM)*.

[54] Jonathan Zittrain and Benjamin Edelman. 2003. Internet Filtering in China. *IEEE Internet Computing* 7, 2 (March 2003), 70–77.