Network Working Group
Internet-Draft
Intended status: Informational
Expires: September 12, 2019

J. Hall
CDT
M. Aaron
CU Boulder
S. Adams
CDT
B. Jones
N. Feamster
Princeton

March 11, 2019

A Survey of Worldwide Censorship Techniques
draft-hall-censorship-tech-07

Abstract

This document describes the technical mechanisms used by censorship regimes around the world to block or impair Internet traffic. It aims to make designers, implementers, and users of Internet protocols aware of the properties being exploited and mechanisms used to censor end-user access to information. This document makes no suggestions on individual protocol considerations, and is purely informational, intended to be a reference.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on September 12, 2019.

Copyright Notice

Copyright (c) 2019 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (https://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction

Censorship is where an entity in a position of power - such as a government, organization, or individual - suppresses communication that it considers objectionable, harmful, sensitive, politically incorrect or inconvenient. (Although censors that engage in censorship must do so through legal, military, or other means, this document focuses largely on technical mechanisms used to achieve network censorship.)

This document describes the technical mechanisms that censorship regimes around the world use to block or degrade Internet traffic (see [RFC7754] for a discussion of Internet blocking and filtering in terms of implications for Internet architecture, rather than end-user access to content and services).

We describe three elements of Internet censorship: prescription, identification, and interference. Prescription is the process by which censors determine what types of material they should block, i.e. they decide to block a list of pornographic websites. Identification is the process by which censors classify specific traffic to be blocked or impaired, i.e. the censor blocks or impairs all webpages containing "sex" in the title or traffic to www.sex.example. Interference is the process by which the censor intercedes in communication and prevents access to censored materials by blocking access or impairing the connection.

2. Technical Prescription

Prescription is the process of figuring out what censors would like to block [Glanville-2008]. Generally, censors aggregate information "to block" in blacklists or using real-time heuristic assessment of content [Ding-1999]. There are indications that online censors are starting to use machine learning techniques as well [Tang-2016].

There are typically three types of blacklists: Keyword, domain name, or Internet Protocol (IP) address. Keyword and domain name blocking take place at the application level (e.g. HTTP), whereas IP blocking tends to take place using routing data in TCP/IP headers. The mechanisms for building up these blacklists are varied. Censors can purchase from private industry "content control" software, such as SmartFilter, which allows filtering from broad categories that they would like to block, such as gambling or pornography. In these cases, these private services attempt to categorize every semi-questionable website as to allow for meta-tag blocking (similarly, they tune real-time content heuristic systems to map their assessments onto categories of objectionable content).

Countries that are more interested in retaining specific political control, a desire which requires swift and decisive action, often have ministries or organizations, such as the Ministry of Industry and Information Technology in China or the Ministry of Culture and Islamic Guidance in Iran, which maintain their own blacklists.

3. Technical Identification

3.1. Points of Control

Internet censorship, necessarily, takes place over a network. Network design gives censors a number of different points-of-control where they can identify the content they are interested in filtering. An important aspect of pervasive technical interception is the necessity to rely on software or hardware to intercept the content the censor is interested in. This requirement, the need to have the interception mechanism located somewhere, logically or physically, implicates various general points-of-control:

- *Internet Backbone:* If a censor controls the gateways into a region, they can filter undesirable traffic that is traveling into and out of the region by packet sniffing and port mirroring at the relevant exchange points. Censorship at this point of control is most effective at controlling the flow of information between a region and the rest of the Internet, but is ineffective at identifying content traveling between the users within a region.

- *Internet Service Providers:* Internet Service Providers are perhaps the most natural point of control. They have a benefit of being easily enumerable by a censor paired with the ability to identify the regional and international traffic of all their users. The censor's filtration mechanisms can be placed on an ISP via governmental mandates, ownership, or voluntary/coercive influence.

- *Institutions:* Private institutions such as corporations, schools, and cyber cafes can put filtration mechanisms in place. These mechanisms are occasionally at the request of a censor, but are more often implemented to help achieve institutional goals, such as to prevent the viewing of pornography on school computers.

- *Personal Devices:* Censors can mandate censorship software be installed on the device level. This has many disadvantages in terms of scalability, ease-of-circumvention, and operating system requirements. The emergence of mobile devices exacerbate these feasibility problems.

- *Services:* Application service providers can be pressured, coerced, or legally required to censor specific content or flows of data. Service providers naturally face incentives to maximize their potential customer base and potential service shutdowns or legal liability due to censorship efforts may seem much less attractive than potentially excluding content, users, or uses of their service.

- *Certificate Authorities:* Authorities that issue cryptographically secured resources can be a significant point of control. Certificate Authorities that issue certificates to domain holders for TLS/HTTPS or Regional/Local Internet Registries that issue Route Origination Authorizations to BGP operators can be forced to issue rogue certificates that may allow compromises in confidentiality guarantees - allowing censorship software to engage in identification and interference where not possible before - or integrity guarantees - allowing, for example, adversarial routing of traffic.

- *Content Distribution Networks (CDNs):* CDNs seek to collapse network topology in order to better locate content closer to the service's users in order to improve quality of service. These can be powerful points of control for censors, especially if the location of a CDN results in easier interference.

At all levels of the network hierarchy, the filtration mechanisms used to detect undesirable traffic are essentially the same: a censor sniffs transmitting packets and identifies undesirable content, and then uses a blocking or shaping mechanism to prevent or impair access. Identification of undesirable traffic can occur at the application, transport, or network layer of the IP stack. Censors are almost always concerned with web traffic, so the relevant protocols tend to be filtered in predictable ways. For example, a subversive image would always make it past a keyword filter, but the IP address of the site serving the image may be blacklisted when identified as a provider of undesirable content.

3.2. Application Layer

3.2.1. HTTP Request Header Identification

An HTTP header contains a lot of useful information for traffic identification; although "host" is the only required field in an HTTP request header (for HTTP/1.1 and later), an HTTP method field is necessary to do anything useful. As such, "method" and "host" are the two fields used most often for ubiquitous censorship. A censor can sniff traffic and identify a specific domain name (host) and usually a page name (GET /page) as well. This identification

technique is usually paired with TCP/IP header identification (see Section 3.3.1) for a more robust method.

Tradeoffs: Request Identification is a technically straight-forward identification method that can be easily implemented at the Backbone or ISP level. The hardware needed for this sort of identification is cheap and easy-to-acquire, making it desirable when budget and scope are a concern. HTTPS will encrypt the relevant request and response fields, so pairing with TCP/IP identification (see Section 3.3.1) is necessary for filtering of HTTPS. However, some countermeasures such as URL obfuscation [RSF-2005] can trivially defeat simple forms of HTTP Request Header Identification.

Empirical Examples: Studies exploring censorship mechanisms have found evidence of HTTP header/ URL filtering in many countries, including Bangladesh, Bahrain, China, India, Iran, Malaysia, Pakistan, Russia, Saudi Arabia, South Korea, Thailand, and Turkey [Verkamp-2012] [Nabi-2013] [Aryan-2012]. Commercial technologies such as the McAfee SmartFilter and NetSweeper are often purchased by censors [Dalek-2013]. These commercial technologies use a combination of HTTP Request Identification and TCP/IP Header Identification to filter specific URLs. Dalek et al. and Jones et al. identified the use of these products in the wild [Dalek-2013] [Jones-2014].

3.2.2. HTTP Response Header Identification

While HTTP Request Header Identification relies on the information contained in the HTTP request from client to server, response identification uses information sent in response by the server to client to identify undesirable content.

Tradeoffs: As with HTTP Request Header Identification, the techniques used to identify HTTP traffic are well-known, cheap, and relatively easy to implement, but is made useless by HTTPS, because the response in HTTPS is encrypted, including headers.

The response fields are also less helpful for identifying content than request fields, as "Server" could easily be identified using HTTP Request Header identification, and "Via" is rarely relevant. HTTP Response censorship mechanisms normally let the first n packets through while the mirrored traffic is being processed; this may allow some content through and the user may be able to detect that the censor is actively interfering with undesirable content.

Empirical Examples: In 2009, Jong Park et al. at the University of New Mexico demonstrated that the Great Firewall of China (GFW) has used this technique [Crandall-2010]. However, Jong Park et al. found

that the GFW discontinued this practice during the course of the study. Due to the overlap in HTTP response filtering and keyword filtering (see Section 3.2.3), it is likely that most censors rely on keyword filtering over TCP streams instead of HTTP response filtering.

3.2.3. Instrumenting Content Providers

In addition to censorship by the state, many governments pressure content providers to censor themselves. Due to the extensive reach of government censorship, we need to define content provider as any service that provides utility to users, including everything from web sites to locally installed programs. The defining factor of keyword identification by content providers is the choice of content providers to detect restricted terms on their platform. The terms to look for may be provided by the government or the content provider may be expected to come up with their own list.

Tradeoffs: By instrumenting content providers to identify restricted content, the censor can gain new information at the cost of political capital with the companies it forces or encourages to participate in censorship. For example, the censor can gain insight about the content of encrypted traffic by coercing web sites to identify restricted content, but this may drive away potential investment. Coercing content providers may encourage self-censorship, an additional advantage for censors. The tradeoffs for instrumenting content providers are highly dependent on the content provider and the requested assistance.

Empirical Examples: Researchers have discovered keyword identification by content providers on platforms ranging from instant messaging applications [Senft-2013] to search engines [Rushe-2015] [Cheng-2010] [Whittaker-2013] [BBC-2013] [Condliffe-2013]. To demonstrate the prevalence of this type of keyword identification, we look to search engine censorship.

Search engine censorship demonstrates keyword identification by content providers and can be regional or worldwide. Implementation is occasionally voluntary, but normally is based on laws and regulations of the country a search engine is operating in. The keyword blacklists are most likely maintained by the search engine provider. China is known to require search engine providers to "voluntarily" maintain search term blacklists to acquire/keep an Internet content provider (ICP) license [Cheng-2010]. It is clear these blacklists are maintained by each search engine provider based on the slight variations in the intercepted searches [Zhu-2011] [Whittaker-2013]. The United Kingdom has been pushing search engines to self-censor with the threat of litigation if they don’t do it

themselves: Google and Microsoft have agreed to block more than 100,000 queries in U.K. to help combat abuse [BBC-2013] [Condliffe-2013].

Depending on the output, search engine keyword identification may be difficult or easy to detect. In some cases specialized or blank results provide a trivial enumeration mechanism, but more subtle censorship can be difficult to detect. In February 2015, Microsoft's search engine, Bing, was accused of censoring Chinese content outside of China [Rushe-2015] because Bing returned different results for censored terms in Chinese and English. However, it is possible that censorship of the largest base of Chinese search users, China, biased Bing's results so that the more popular results in China (the uncensored results) were also more popular for Chinese speakers outside of China.

3.2.4. Deep Packet Inspection (DPI) Identification

Deep Packet Inspection has become computationally feasible as a censorship mechanism in recent years [Wagner-2009]. Unlike other techniques, DPI reassembles network flows to examine the application "data" section, as opposed to only the header, and is therefore often used for keyword identification. DPI also differs from other identification technologies because it can leverage additional packet and flow characteristics, i.e. packet sizes and timings, to identify content. To prevent substantial quality of service (QoS) impacts, DPI normally analyzes a copy of data while the original packets continue to be routed. Typically, the traffic is split using either a mirror switch or fiber splitter, and analyzed on a cluster of machines running Intrusion Detection Systems (IDS) configured for censorship.

Tradeoffs: DPI is one of the most expensive identification mechanisms and can have a large QoS impact [Porter-2010]. When used as a keyword filter for TCP flows, DPI systems can cause also major overblocking problems. Like other techniques, DPI is less useful against encrypted data, though DPI can leverage unencrypted elements of an encrypted data flow (e.g., the Server Name Indicator (SNI) sent in the clear for TLS) or statistical information about an encrypted flow (e.g., video takes more bandwidth than audio or textual forms of communication) to identify traffic.

Other kinds of information can be inferred by comparing certain unencrypted elements exchanged during TLS handshakes to similar data points from known sources. This practice, called TLS fingerprinting, allows a probabilistic identification of a party's operating system, browser, or application based on a comparison of the specific combinations of TLS version, ciphersuites, compression options, etc.

sent in the ClientHello message to similar signatures found in unencrypted traffic [Husak-2016].

Despite these problems, DPI is the most powerful identification method and is widely used in practice. The Great Firewall of China (GFW), the largest censorship system in the world, has used DPI to identify restricted content over HTTP and DNS and inject TCP RSTs and bad DNS responses, respectively, into connections [Crandall-2010] [Clayton-2006] [Anonymous-2014].

Empirical Examples: Several studies have found evidence of DPI being used to censor content and tools. Clayton et al. Crandal et al., Anonymous, and Khattak et al., all explored the GFW and Khattak et al. even probed the firewall to discover implementation details like how much state it stores [Crandall-2010] [Clayton-2006] [Anonymous-2014] [Khattak-2013]. The Tor project claims that China, Iran, Ethiopia, and others must have used DPI to block the obsf2 protocol [Wilde-2012]. Malaysia has been accused of using targeted DPI, paired with DDoS, to identify and subsequently knockout pro-opposition material [Wagstaff-2013]. It also seems likely that organizations not so worried about blocking content in real-time could use DPI to sort and categorically search gathered traffic using technologies such as NarusInsight [Hepting-2011].

3.2.4.1. Server Name Indication

In encrypted connections using Transport Layer Security (TLS), there may be servers that host multiple "virtual servers" at a give network address, and the client will need to specify in the (unencrypted) Client Hello message which domain name it seeks to connect to (so that the server can respond with the appropriate TLS certificate) using the Server Name Indication (SNI) TLS extension [RFC6066]. Since SNI is sent in the clear, censors and filtering software can use it as a basis for blocking, filtering, or impairment by dropping connections to domains that match prohibited content (e.g., bad.foo.example may be censored while good.foo.example is not) [Shbair-2015].

Domain fronting has been one popular way to avoid identification by censors [Fifield-2015]. To avoid identification by censors, applications using domain fronting put a different domain name in the SNI extension than the one encrypted by HTTPS. The visible SNI would indicate an unblocked domain, while the blocked domain remains hidden in the encrypted application header. Some encrypted messaging services relied on domain fronting to enable their provision in countries employing SNI-based filtering. These services used the cover provided by domains for which blocking at the domain level would be undesirable to hide their true domain names. However, the

companies holding the most popular domains have since reconfigured their software to prevent this practice. It may be possible to achieve similar results using potential future options to encrypt SNI in TLS 1.3.

Tradeoffs: Some clients do not send the SNI extension (e.g., clients that only support versions of SSL and not TLS) or will fall back to SSL if a TLS connection fails, rendering this method ineffective. In addition, this technique requires deep packet inspection techniques that can be computationally and infrastructurally expensive and improper configuration of an SNI-based block can result in significant overblocking, e.g., when a second-level domain like populardomain.example is inadvertently blocked. In the case of encrypted SNI, pressure to censor may transfer to other points of intervention, such as content and application providers.

Empirical Examples: While there are many examples of security firms that offer SNI-based filtering [Trustwave-2015] [Sophos-2015] [Shbair-2015], the government of South Korea was recently observed using SNI-based filtering. Cite to Gatlan https://www.bleepingcomputer.com/news/security/south-korea-is-censoring-the-internet-by-snooping-on-sni-traffic/

3.3. Transport Layer

3.3.1. Shallow Packet Inspection and TCP/IP Header Identification

Of the various shallow packet inspection methods, TCP/IP Header Identification is the most pervasive, reliable, and predictable type of identification. TCP/IP headers contain a few invaluable pieces of information that must be transparent for traffic to be successfully routed: destination and source IP address and port. Destination and Source IP are doubly useful, as not only does it allow a censor to block undesirable content via IP blacklisting, but also allows a censor to identify the IP of the user making the request. Port is useful for whitelisting certain applications.

Trade-offs: TCP/IP identification is popular due to its simplicity, availability, and robustness.

TCP/IP identification is trivial to implement, but is difficult to implement in backbone or ISP routers at scale, and is therefore typically implemented with DPI. Blacklisting an IP is equivalent to installing a /32 route on a router and due to limited flow table space, this cannot scale beyond a few thousand IPs at most. IP blocking is also relatively crude, leading to overblocking, and cannot deal with some services like Content Distribution Networks

(CDN), that host content at hundreds or thousands of IP addresses. Despite these limitations, IP blocking is extremely effective because the user needs to proxy their traffic through another destination to circumvent this type of identification.

Port-blocking is generally not useful because many types of content share the same port and it is possible for censored applications to change their port. For example, most HTTP traffic goes over port 80, so the censor cannot differentiate between restricted and allowed content solely on the basis of port. Port whitelisting is occasionally used, where a censor limits communication to approved ports, such as 80 for HTTP traffic and is most effective when used in conjunction with other identification mechanisms. For example, a censor could block the default HTTPS port, port 443, thereby forcing most users to fall back to HTTP.

3.3.2. Protocol Identification

Censors sometimes identify entire protocols to be blocked using a variety of traffic characteristics. For example, Iran impairs the performance of HTTPS traffic, a protocol that prevents further analysis, to encourage users to switch to HTTP, a protocol that they can analyze [Aryan-2012]. A simple protocol identification would be to recognize all TCP traffic over port 443 as HTTPS, but more sophisticated analysis of the statistical properties of payload data and flow behavior, would be more effective, even when port 443 is not used [Hjelmvik-2010] [Sandvine-2014].

If censors can detect circumvention tools, they can block them, so censors like China are extremely interested in identifying the protocols for censorship circumvention tools. In recent years, this has devolved into an arms race between censors and circumvention tool developers. As part of this arms race, China developed an extremely effective protocol identification technique that researchers call active probing or active scanning.

In active probing, the censor determines whether hosts are running a circumvention protocol by trying to initiate communication using the circumvention protocol. If the host and the censor successfully negotiate a connection, then the censor conclusively knows that host is running a circumvention tool. China has used active scanning to great effect to block Tor [Winter-2012].

Trade-offs: Protocol Identification necessarily only provides insight into the way information is traveling, and not the information itself.

Protocol identification is useful for detecting and blocking circumvention tools, like Tor, or traffic that is difficult to analyze, like VoIP or SSL, because the censor can assume that this traffic should be blocked. However, this can lead to over-blocking problems when used with popular protocols. These methods are expensive, both computationally and financially, due to the use of statistical analysis, and can be ineffective due to its imprecise nature.

Empirical Examples: Protocol identification can be easy to detect if it is conducted in real time and only a particular protocol is blocked, but some types of protocol identification, like active scanning, are much more difficult to detect. Protocol identification has been used by Iran to identify and throttle SSH traffic to make it unusable [Anonymous-2007] and by China to identify and block Tor relays [Winter-2012]. Protocol Identification has also been used for traffic management, such as the 2007 case where Comcast in the United States used RST injection to interrupt BitTorrent Traffic [Winter-2012].

4. Technical Interference

4.1. Application Layer

4.1.1. DNS Interference

There are a variety of mechanisms that censors can use to block or filter access to content by altering responses from the DNS [AFNIC-2013] [ICANN-SSAC-2012], including blocking the response, replying with an error message, or responding with an incorrect address.

"DNS mangling" is a network-level technique where an incorrect IP address is returned in response to a DNS query to a censored destination. An example of this is what some Chinese networks do (we are not aware of any other wide-scale uses of mangling). On those Chinese networks, every DNS request in transit is examined (presumably by network inspection technologies such as DPI) and, if it matches a censored domain, a false response is injected. End users can see this technique in action by simply sending DNS requests to any unused IP address in China (see example below). If it is not a censored name, there will be no response. If it is censored, an erroneous response will be returned. For example, using the command-line dig utility to query an unused IP address in China of 192.0.2.2 for the name "www.uncensored.example" compared with "www.censored.example" (censored at the time of writing), we get an erroneous IP address "198.51.100.0" as a response:

```
% dig +short +nodnssec @192.0.2.2 A www.uncensored.example
;; connection timed out; no servers could be reached

% dig +short +nodnssec @192.0.2.2 A www.censored.example
198.51.100.0
```

There are also cases of what is colloquially called "DNS lying", where a censor mandates that the DNS responses provided - by an operator of a recursive resolver such as an Internet access provider - be different than what authoritative resolvers would provide [Bortzmayer-2015].

DNS cache poisoning refers to a mechanism where a censor interferes with the response sent by an authoritative DNS resolver to a recursive resolver by responding more quickly than the authoritative resolver can respond with an alternative IP address [Halley-2008]. Cache poisoning occurs after the requested site's name servers resolve the request and attempt to forward the true IP back to the requesting device; on the return route the resolved IP is recursively cached by each DNS server that initially forwarded the request. During this caching process if an undesirable keyword is recognized, the resolved IP is "poisoned" and an alternative IP (or NXDOMAIN error) is returned more quickly than the upstream resolver can respond, causing an erroneous IP address to be cached (and potentially recursively so). The alternative IPs usually direct to a nonsense domain or a warning page. Alternatively, Iranian censorship appears to prevent the communication en-route, preventing a response from ever being sent [Aryan-2012].

Trade-offs: These forms of DNS interference require the censor to force a user to traverse a controlled DNS hierarchy (or intervening network on which the censor serves as a Active Pervasive Attacker [RFC7624] to rewrite DNS responses) for the mechanism to be effective. It can be circumvented by a technical savvy user that opts to use alternative DNS resolvers (such as the public DNS resolvers provided by Google, OpenDNS, Telcomix, or FDN) or Virtual Private Network technology. DNS mangling and cache poisoning also imply returning an incorrect IP to those attempting to resolve a domain name, but in some cases the destination may be technically accessible; over HTTP, for example, the user may have another method of obtaining the IP address of the desired site and may be able to access it if the site is configured to be the default server listening at this IP address. Target blocking has also been a problem, as occasionally users outside of the censors region will be directed through DNS servers or DNS-rewriting network equipment controlled by a censor, causing the request to fail. The ease of circumvention paired with the large risk of content blocking and target blocking make DNS interference a partial, difficult, and less

than ideal censorship mechanism. Additionally, the above mechanisms rely on DNSSEC not being deployed or DNSSEC validation not being active on the client or recursive resolver.

Empirical Examples: DNS interference, when properly implemented, is easy to identify based on the shortcomings identified above. Turkey relied on DNS interference for its country-wide block of websites such Twitter and YouTube for almost week in March of 2014 but the ease of circumvention resulted in an increase in the popularity of Twitter until Turkish ISPs implementing an IP blacklist to achieve the governmental mandate [Zmijewki-2014]. Ultimately, Turkish ISPs started hijacking all requests to Google and Level 3's international DNS resolvers [Zmijewki-2014]. DNS interference, when incorrectly implemented, has resulted in some of the largest "censorship disasters". In January 2014, China started directing all requests passing through the Great Fire Wall to a single domain, dongtaiwang.com, due to an improperly configured DNS poisoning attempt; this incident is thought to be the largest Internet-service outage in history [AFP-2014] [Anon-SIGCOMM12]. Countries such as China, Iran, Turkey, and the United States have discussed blocking entire TLDs as well, but only Iran has acted by blocking all Israeli (.il) domains [Albert-2011].

4.2. Transport Layer

4.2.1. Performance Degradation

While other interference techniques outlined in this section mostly focus on blocking or preventing access to content, it can be an effective censorship strategy in some cases to not entirely block access to a given destination, or service but instead degrade the performance of the relevant network connection. The resulting user experience for a site or service under performance degradation can be so bad that users opt to use a different site, service, or method of communication, or may not engage in communication at all if there are no alternatives. Traffic shaping techniques that rate-limit the bandwidth available to certain types of traffic is one example of a performance degradation.

Trade offs: While implementing a performance degradation will not always eliminate the ability of people to access a desire resource, it may force them to use other means of communication where censorship (or surveillance) is more easily accomplished.

Empirical Examples: Iran has been known to shape the bandwidth available to HTTPS traffic to encourage unencrypted HTTP traffic [Aryan-2012].

4.2.2. Packet Dropping

Packet dropping is a simple mechanism to prevent undesirable traffic. The censor identifies undesirable traffic and chooses to not properly forward any packets it sees associated with the traversing undesirable traffic instead of following a normal routing protocol. This can be paired with any of the previously described mechanisms so long as the censor knows the user must route traffic through a controlled router.

Trade offs: Packet Dropping is most successful when every traversing packet has transparent information linked to undesirable content, such as a Destination IP. One downside Packet Dropping suffers from is the necessity of blocking all content from otherwise allowable IPs based on a single subversive sub-domain; blogging services and github repositories are good examples. China famously dropped all github packets for three days based on a single repository hosting undesirable content [Anonymous-2013]. The need to inspect every traversing packet in close to real time also makes Packet Dropping somewhat challenging from a QoS perspective.

Empirical Examples: Packet Dropping is a very common form of technical interference and lends itself to accurate detection given the unique nature of the time-out requests it leaves in its wake. The Great Firewall of China has been observed using packet dropping as one of its primary mechanisms of technical censorship [Ensafi-2013]. Iran has also used Packet Dropping as the mechanisms for throttling SSH [Aryan-2012]. These are but two examples of a ubiquitous censorship practice.

4.2.3. RST Packet Injection

Packet injection, generally, refers to a man-in-the-middle (MITM) network interference technique that spoofs packets in an established traffic stream. RST packets are normally used to let one side of TCP connection know the other side has stopped sending information, and thus the receiver should close the connection. RST Packet Injection is a specific type of packet injection attack that is used to interrupt an established stream by sending RST packets to both sides of a TCP connection; as each receiver thinks the other has dropped the connection, the session is terminated.

Trade-offs: RST Packet Injection has a few advantages that make it extremely popular as a censorship technique. RST Packet Injection is an out-of-band interference mechanism, allowing the avoidance of the the QoS bottleneck one can encounter with inline techniques such as Packet Dropping. This out-of-band property allows a censor to inspect a copy of the information, usually mirrored by an optical

splitter, making it an ideal pairing for DPI and Protocol Identification [Weaver-2009] (this asynchronous version of a MITM is often called a Man-on-the-Side (MOTS)). RST Packet Injection also has the advantage of only requiring one of the two endpoints to accept the spoofed packet for the connection to be interrupted.

The difficult part of RST Packet Injection is spoofing "enough" correct information to ensure one end-point accepts a RST packet as legitimate; this generally implies a correct IP, port, and (TCP) sequence number. Sequence number is the hardest to get correct, as [RFC0793] specifies an RST Packet should be in-sequence to be accepted, although the RFC also recommends allowing in-window packets as "good enough". This in-window recommendation is important, as if it is implemented it allows for successful Blind RST Injection attacks [Netsec-2011]. When in-window sequencing is allowed, It is trivial to conduct a Blind RST Injection, a blind injection implies the censor doesn't know any sensitive (encrypted) sequencing information about the TCP stream they are injecting into, they can simply enumerate the ~70000 possible windows; this is particularly useful for interrupting encrypted/obfuscated protocols such as SSH or Tor. RST Packet Injection relies on a stateful network, making it useless against UDP connections. RST Packet Injection is among the most popular censorship techniques used today given its versatile nature and effectiveness against all types of TCP traffic.

Empirical Examples: RST Packet Injection, as mentioned above, is most often paired with identification techniques that require splitting, such as DPI or Protocol Identification. In 2007, Comcast was accused of using RST Packet Injection to interrupt traffic it identified as BitTorrent [Schoen-2007], this later led to a US Federal Communications Commission ruling against Comcast [VonLohmann-2008]. China has also been known to use RST Packet Injection for censorship purposes. This interference is especially evident in the interruption of encrypted/obfuscated protocols, such as those used by Tor [Winter-2012].

4.3. Multi-layer and Non-layer

4.3.1. Distributed Denial of Service (DDoS)

Distributed Denial of Service attacks are a common attack mechanism used by "hacktivists" and malicious hackers, but censors have used DDoS in the past for a variety of reasons. There is a huge variety of DDoS attacks [Wikip-DoS], but on a high level two possible impacts tend to occur; a flood attack results in the service being unusable while resources are being spent to flood the service, a crash attack aims to crash the service so resources can be reallocated elsewhere without "releasing" the service.

Trade-offs: DDoS is an appealing mechanism when a censor would like to prevent all access to undesirable content, instead of only access in their region for a limited period of time, but this is really the only uniquely beneficial feature for DDoS as a censorship technique. The resources required to carry out a successful DDoS against major targets are computationally expensive, usually requiring renting or owning a malicious distributed platform such as a botnet, and imprecise. DDoS is an incredibly crude censorship technique, and appears to largely be used as a timely, easy-to-access mechanism for blocking undesirable content for a limited period of time.

Empirical Examples: In 2012 the U.K.'s GCHQ used DDoS to temporarily shutdown IRC chat rooms frequented by members of Anonymous using the Syn Flood DDoS method; Syn Flood exploits the handshake used by TCP to overload the victim server with so many requests that legitimate traffic becomes slow or impossible [Schone-2014] [CERT-2000]. Dissenting opinion websites are frequently victims of DDoS around politically sensitive events in Burma [Villeneuve-2011]. Controlling parties in Russia [Kravtsova-2012], Zimbabwe [Orion-2013], and Malaysia [Muncaster-2013] have been accused of using DDoS to interrupt opposition support and access during elections. In 2015, China launched a DDoS attack using a true MITM system collocated with the Great Firewall, dubbed "Great Cannon", that was able to inject JavaScript code into web visits to a Chinese search engine that commandeered those user agents to send DDoS traffic to various sites [Marczak-2015].

4.3.2. Network Disconnection or Adversarial Route Announcement

While it is perhaps the crudest of all censorship techniques, there is no more effective way of making sure undesirable information isn't allowed to propagate on the web than by shutting off the network. The network can be logically cut off in a region when a censoring body withdraws all of the Boarder Gateway Protocol (BGP) prefixes routing through the censor's country.

Trade-offs: The impact to a network disconnection in a region is huge and absolute; the censor pays for absolute control over digital information with all the benefits the Internet brings; this is never a long-term solution for any rational censor and is normally only used as a last resort in times of substantial unrest.

Empirical Examples: Network Disconnections tend to only happen in times of substantial unrest, largely due to the huge social, political, and economic impact such a move has. One of the first, highly covered occurrences was with the Junta in Myanmar employing Network Disconnection to help Junta forces quash a rebellion in 2007

[Dobie-2007]. China disconnected the network in the Xinjiang region during unrest in 2009 in an effort to prevent the protests from spreading to other regions [Heacock-2009]. The Arab Spring saw the the most frequent usage of Network Disconnection, with events in Egypt and Libya in 2011 [Cowie-2011] [Cowie-2011b], and Syria in 2012 [Thomson-2012]. Russia has indicated that it will attempt to disconnect all Russian networks from the global internet in April 2019 as part of a test of the nation's network independence. Reports also indicate that, as part of the test disconnect, Russian telecom firms must route all traffic to state-operated monitoring points. cite ZD Net https://www.zdnet.com/article/russia-to-disconnect-from-the-internet-as-part-of-a-planned-test/

5. Non-Technical Prescription

As the name implies, sometimes manpower is the easiest way to figure out which content to block. Manual Filtering differs from the common tactic of building up blacklists in that it doesn't necessarily target a specific IP or DNS, but instead removes or flags content. Given the imprecise nature of automatic filtering, manually sorting through content and flagging dissenting websites, blogs, articles and other media for filtration can be an effective technique. This filtration can occur on the Backbone/ISP level - China's army of monitors is a good example [BBC-2013b] - but more commonly manual filtering occurs on an institutional level. Internet Content Providers such as Google or Weibo, require a business license to operate in China. One of the prerequisites for a business license is an agreement to sign a "voluntary pledge" known as the "Public Pledge on Self-discipline for the Chinese Internet Industry". The failure to "energetically uphold" the pledged values can lead to the ICPs being held liable for the offending content by the Chinese government [BBC-2013b].

6. Non-Technical Interference

6.1. Self-Censorship

Self-censorship is one of the most interesting and effective types of censorship; a mix of Bentham's Panopticon, cultural manipulation, intelligence gathering, and meatspace enforcement. Simply put, self-censorship is when a censor creates an atmosphere where users censor themselves. This can be achieved through controlling information, intimidating would-be dissidents, swaying public thought, and creating apathy. Self-censorship is difficult to document, as when it is implemented effectively the only noticeable tracing is a lack of undesirable content; instead one must look at the tools and techniques used by censors to encourage self-censorship. Controlling Information relies on traditional censorship techniques, or by

forcing all users to connect through an intranet, such as in North Korea. Intimidation is often achieved through allowing Internet users to post "whatever they want," but arresting those who post about dissenting views, this technique is incredibly common [Calamur-2013] [AP-2012] [Hopkins-2011] [Guardian-2014] [Johnson-2010]. A good example of swaying public thought is China's "50-Cent Party," reported to be composed of somewhere between 20,000 [Bristow-2013] and 300,000 [Fareed-2008] contributors who are paid to "guide public thought" on local and regional issues as directed by the Ministry of Culture. Creating apathy can be a side-effect of successfully controlling information over time and is ideal for a censorship regime [Gao-2014].

6.2. Domain Name Reallocation

Because domain names are resolved recursively, if a root name server reassigns or delists a domain, all other DNS servers will be unable to properly forward and cache the site. Domain name registration is only really a risk where undesirable content is hosted on TLD controlled by the censoring country, such as .cn or .ru [Anderson-2011] or where legal processes in countries like the United States result in domain name seizures and/or DNS redirection by the government [Kopel-2013].

6.3. Server Takedown

Servers must have a physical location somewhere in the world. If undesirable content is hosted in the censoring country the servers can be physically seized or the hosting provider can be required to prevent access [Anderson-2011].

6.4. Notice and Takedown

In some countries, legal mechanisms exist where an individual can issue a legal request to a content host that requires the host to take down content. Examples include the voluntary systems employed by companies like Google to comply with "Right to be Forgotten" policies in the European Union [Google-RTBF] and the copyright-oriented notice and takedown regime of the United States Digital Millennium Copyright Act (DMCA) Section 512 [DMLP-512].

7. Contributors

This document benefited from discussions with Stephane Bortzmeyer, Nick Feamster, and Martin Nilsson.

8. Informative References

[AFNIC-2013]
AFNIC, "Report of the AFNIC Scientific Council: Consequences of DNS-based Internet filtering", 2013, <http://www.afnic.fr/medias/documents/conseilscientifique/SC-consequences-of-DNS-based-Internet-filtering.pdf>.

[AFP-2014]
AFP, "China Has Massive Internet Breakdown Reportedly Caused By Their Own Censoring Tools", 2014, <http://www.businessinsider.com/chinas-internet-breakdown-reportedly-caused-by-censoring-tools-2014-1>.

[Albert-2011]
Albert, K., "DNS Tampering and the new ICANN gTLD Rules", 2011, <https://opennet.net/blog/2011/06/dns-tampering-and-new-icann-gtld-rules>.

[Anderson-2011]
Anderson, R. and S. Murdoch, "Access Denied: Tools and Technology of Internet Filtering", 2011, <http://access.opennet.net/wp-content/uploads/2011/12/accessdenied-chapter-3.pdf>.

[Anon-SIGCOMM12]
Anonymous, "The Collateral Damage of Internet Censorship by DNS Injection", 2012, <http://www.sigcomm.org/sites/default/files/ccr/papers/2012/July/2317307-2317311.pdf>.

[Anonymous-2007]
Anonymous, "How to Bypass Comcast's Bittorrent Throttling", 2012, <https://torrentfreak.com/how-to-bypass-comcast-bittorrent-throttling-071021>.

[Anonymous-2013]
Anonymous, "GitHub blocked in China - how it happened, how to get around it, and where it will take us", 2013, <https://en.greatfire.org/blog/2013/jan/github-blocked-china-how-it-happened-how-get-around-it-and-where-it-will-take-us>.

[Anonymous-2014]
Anonymous, "Towards a Comprehensive Picture of the Great Firewall's DNS Censorship", 2014, <https://www.usenix.org/system/files/conference/foci14/foci14-anonymous.pdf>.

[AP-2012] Associated Press, "Sattar Beheshit, Iranian Blogger, Was Beaten In Prison According To Prosecutor", 2012, <http://www.huffingtonpost.com/2012/12/03/sattar-beheshit-iran_n_2233125.html>.

[Aryan-2012] Aryan, S., Aryan, H., and J. Halderman, "Internet Censorship in Iran: A First Look", 2012, <https://jhalderm.com/pub/papers/iran-foci13.pdf>.

[BBC-2013] BBC News, "Google and Microsoft agree steps to block abuse images", 2013, <http://www.bbc.com/news/uk-24980765>.

[BBC-2013b] BBC, "China employs two million microblog monitors state media say", 2013, <http://www.bbc.com/news/world-asia-china-2439695>.

[Bortzmayer-2015] Bortzmayer, S., "DNS Censorship (DNS Lies) As Seen By RIPE Atlas", 2015, <https://labs.ripe.net/Members/stephane_bortzmeyer/dns-censorship-dns-lies-seen-by-atlas-probes>.

[Bristow-2013] Bristow, M., "China's internet 'spin doctors'", 2013, <http://news.bbc.co.uk/2/hi/asia-pacific/7783640.stm>.

[Calamur-2013] Calamur, K., "Prominent Egyptian Blogger Arrested", 2013, <http://www.npr.org/blogs/thetwo-way/2013/11/29/247820503/prominent-egyptian-blogger-arrested>.

[CERT-2000] CERT, "TCP SYN Flooding and IP Spoofing Attacks", 2000, <http://www.cert.org/historical/advisories/CA-1996-21.cfm>.

[Cheng-2010] Cheng, J., "Google stops Hong Kong auto-redirect as China plays hardball", 2010, <http://arstechnica.com/tech-policy/2010/06/google-tweaks-china-to-hong-kong-redirect-same-results/>.

[Clayton-2006] Clayton, R., "Ignoring the Great Firewall of China", 2006, <http://link.springer.com/chapter/10.1007/11957454_2>.

[Condliffe-2013]
Condliffe, J., "Google Announces Massive New Restrictions on Child Abuse Search Terms", 2013, <http://gizmodo.com/google-announces-massive-new-restrictions-on-child-abus-1466539163>.

[Cowie-2011]
Cowie, J., "Egypt Leaves the Internet", 2011, <http://www.renesys.com/2011/01/egypt-leaves-the-internet/>.

[Cowie-2011b]
Cowie, J., "Libyan Disconnect", 2011, <http://www.renesys.com/2011/02/libyan-disconnect-1/>.

[Crandall-2010]
Crandall, J., "Empirical Study of a National-Scale Distributed Intrusion Detection System: Backbone-Level Filtering of HTML Responses in China", 2010, <http://www.cs.unm.edu/~crandall/icdcs2010.pdf>.

[Dalek-2013]
Dalek, J., "A Method for Identifying and Confirming the Use of URL Filtering Products for Censorship", 2013, <http://www.cs.stonybrook.edu/~phillipa/papers/imc112s-dalek.pdf>.

[Ding-1999]
Ding, C., Chi, C., Deng, J., and C. Dong, "Centralized Content-Based Web Filtering and Blocking: How Far Can It Go?", 1999, <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.132.3302&rep=rep1&type=pdf>.

[DMLP-512]
Digital Media Law Project, "Protecting Yourself Against Copyright Claims Based on User Content", 2012, <http://www.dmlp.org/legal-guide/protecting-yourself-against-copyright-claims-based-user-content>.

[Dobie-2007]
Dobie, M., "Junta tightens media screw", 2007, <http://news.bbc.co.uk/2/hi/asia-pacific/7016238.stm>.

[Ensafi-2013]
Ensafi, R., "Detecting Intentional Packet Drops on the Internet via TCP/IP Side Channels", 2013, <http://arxiv.org/pdf/1312.5739v1.pdf>.

[Fareed-2008]
Fareed, M., "China joins a turf war", 2008, <http://www.theguardian.com/media/2008/sep/22/chinathemedia.marketingandpr>.

[Fifield-2015]
Fifield, D., Lan, C., Hynes, R., Wegmann, P., and V. Paxson, "Blocking-resistant communication through domain fronting", 2015, <https://petsymposium.org/2015/papers/03_Fifield.pdf>.

[Gao-2014]
Gao, H., "Tiananmen, Forgotten", 2014, <http://www.nytimes.com/2014/06/04/opinion/tiananmen-forgotten.html>.

[Glanville-2008]
Glanville, J., "The Big Business of Net Censorship", 2008, <http://www.theguardian.com/commentisfree/2008/nov/17/censorship-internet>.

[Google-RTBF]
Google, Inc., "Search removal request under data protection law in Europe", 2015, <https://support.google.com/legal/contact/lr_eudpa?product=websearch>.

[Guardian-2014]
The Gaurdian, "Chinese blogger jailed under crackdown on 'internet rumours'", 2014, <http://www.theguardian.com/world/2014/apr/17/chinese-blogger-jailed-crackdown-internet-rumours-qin-zhihui>.

[Halley-2008]
Halley, B., "How DNS cache poisoning works", 2014, <https://www.networkworld.com/article/2277316/tech-primers/tech-primers-how-dns-cache-poisoning-works.html>.

[Heacock-2009]
Heacock, R., "China Shuts Down Internet in Xinjiang Region After Riots", 2009, <https://opennet.net/blog/2009/07/china-shuts-down-internet-xinjiang-region-after-riots>.

[Hepting-2011]
Electronic Frontier Foundation, "Hepting vs. AT&T", 2011, <https://www.eff.org/cases/hepting>.

[Hjelmvik-2010]
Hjelmvik, E., "Breaking and Improving Protocol Obfuscation", 2010, <https://www.iis.se/docs/hjelmvik_breaking.pdf>.

[Hopkins-2011]
Hopkins, C., "Communications Blocked in Libya, Qatari Blogger Arrested: This Week in Online Tyranny", 2011, <http://readwrite.com/2011/03/03/communications_blocked_in_libya_this_week_in_onlin>.

[Husak-2016]
Husak, M., Cermak, M., Jirsik, T., and P. Celeda, "HTTPS traffic analysis and client identification using passive SSL/TLS fingerprinting", 2016, <https://link.springer.com/article/10.1186/s13635-016-0030-7>.

[ICANN-SSAC-2012]
ICANN Security and Stability Advisory Committee (SSAC), "SAC 056: SSAC Advisory on Impacts of Content Blocking via the Domain Name System", 2012, <https://www.icann.org/en/system/files/files/sac-056-en.pdf>.

[Johnson-2010]
Johnson, L., "Torture feared in arrest of Iraqi blogger", 2011, <http://seattlepostglobe.org/2010/02/05/torture-feared-in-arrest-of-iraqi-blogger/>.

[Jones-2014]
Jones, B., "Automated Detection and Fingerprinting of Censorship Block Pages", 2014, <http://conferences2.sigcomm.org/imc/2014/papers/p299.pdf>.

[Khattak-2013]
Khattak, S., "Towards Illuminating a Censorship Monitor's Model to Facilitate Evasion", 2013, <http://0b4af6cdc2f0c5998459-c0245c5c937c5dedcca3f1764ecc9b2f.r43.cf2.rackcdn.com/12389-foci13-khattak.pdf>.

[Kopel-2013]
Kopel, K., "Operation Seizing Our Sites: How the Federal Government is Taking Domain Names Without Prior Notice", 2013, <http://dx.doi.org/doi:10.15779/Z384Q3M>.

[Kravtsova-2012]
Kravtsova, Y., "Cyberattacks Disrupt Opposition's Election", 2012, <http://www.themoscowtimes.com/news/article/cyberattacks-disrupt-oppositions-election/470119.html>.

[Marczak-2015]
Marczak, B., Weaver, N., Dalek, J., Ensafi, R., Fifield, D., McKune, S., Rey, A., Scott-Railton, J., Deibert, R., and V. Paxson, "An Analysis of China's "Great Cannon"", 2015, <https://www.usenix.org/system/files/conference/foci15/foci15-paper-marczak.pdf>.

[Muncaster-2013]
Muncaster, P., "Malaysian election sparks web blocking/ DDoS claims", 2013, <http://www.theregister.co.uk/2013/05/09/malaysia_fraud_elections_ddos_web_blocking/>.

[Nabi-2013]
Nabi, Z., "The Anatomy of Web Censorship in Pakistan", 2013, <http://0b4af6cdc2f0c5998459-c0245c5c937c5dedcca3f1764ecc9b2f.r43.cf2.rackcdn.com/12387-foci13-nabi.pdf>.

[Netsec-2011]
n3t2.3c, "TCP-RST Injection", 2011, <https://nets.ec/TCP-RST_Injection>.

[Orion-2013]
Orion, E., "Zimbabwe election hit by hacking and DDoS attacks", 2013, <http://www.theinquirer.net/inquirer/news/2287433/zimbabwe-election-hit-by-hacking-and-ddos-attacks>.

[Porter-2010]
Porter, T., "The Perils of Deep Packet Inspection", 2010, <http://www.symantec.com/connect/articles/perils-deep-packet-inspection>.

[RFC0793] Postel, J., "Transmission Control Protocol", STD 7, RFC 793, DOI 10.17487/RFC0793, September 1981, <https://www.rfc-editor.org/info/rfc793>.

[RFC6066] Eastlake 3rd, D., "Transport Layer Security (TLS) Extensions: Extension Definitions", RFC 6066, DOI 10.17487/RFC6066, January 2011, <https://www.rfc-editor.org/info/rfc6066>.

[RFC7624] Barnes, R., Schneier, B., Jennings, C., Hardie, T., Trammell, B., Huitema, C., and D. Borkmann, "Confidentiality in the Face of Pervasive Surveillance: A Threat Model and Problem Statement", RFC 7624, DOI 10.17487/RFC7624, August 2015, <https://www.rfc-editor.org/info/rfc7624>.

[RFC7754] Barnes, R., Cooper, A., Kolkman, O., Thaler, D., and E. Nordmark, "Technical Considerations for Internet Service Blocking and Filtering", RFC 7754, DOI 10.17487/RFC7754, March 2016, <https://www.rfc-editor.org/info/rfc7754>.

[RSF-2005] Reporters Sans Frontieres, "Technical ways to get around censorship", 2005, <http://archives.rsf.org/print-blogs.php3?id_article=15013>.

[Rushe-2015] Rushe, D., "Bing censoring Chinese language search results for users in the US", 2013, <http://www.theguardian.com/technology/2014/feb/11/bing-censors-chinese-language-search-results>.

[Sandvine-2014] Sandvine, "Technology Showcase on Traffic Classification: Why Measurements and Freeform Policy Matter", 2014, <https://www.sandvine.com/downloads/general/technology/sandvine-technology-showcases/sandvine-technology-showcase-traffic-classification.pdf>.

[Schoen-2007] Schoen, S., "EFF tests agree with AP: Comcast is forging packets to interfere with user traffic", 2007, <https://www.eff.org/deeplinks/2007/10/eff-tests-agree-ap-comcast-forging-packets-to-interfere>.

[Schone-2014] Schone, M., Esposito, R., Cole, M., and G. Greenwald, "Snowden Docs Show UK Spies Attacked Anonymous, Hackers", 2014, <http://www.nbcnews.com/feature/edward-snowden-interview/exclusive-snowden-docs-show-uk-spies-attacked-anonymous-hackers-n21361>.

[Senft-2013]
Senft, A., "Asia Chats: Analyzing Information Controls and Privacy in Asian Messaging Applications", 2013, <https://citizenlab.org/2013/11/asia-chats-analyzing-information-controls-privacy-asian-messaging-applications/>.

[Shbair-2015]
Shbair, W., Cholez, T., Goichot, A., and I. Chrisment, "Efficiently Bypassing SNI-based HTTPS Filtering", 2015, <https://hal.inria.fr/hal-01202712/document>.

[Sophos-2015]
Sophos, "Understanding Sophos Web Filtering", 2015, <https://www.sophos.com/en-us/support/knowledgebase/115865.aspx>.

[Tang-2016]
Tang, C., "In-depth analysis of the Great Firewall of China", 2016, <https://www.cs.tufts.edu/comp/116/archive/fall2016/ctang.pdf>.

[Thomson-2012]
Thomson, I., "Syria Cuts off Internet and Mobile Communication", 2012, <http://www.theregister.co.uk/2012/11/29/syria_internet_blackout/>.

[Trustwave-2015]
Trustwave, "Filter: SNI extension feature and HTTPS blocking", 2015, <https://www3.trustwave.com/software/8e6/hlp/r3000/files/1system_filter.html>.

[Verkamp-2012]
Verkamp, J. and M. Gupta, "Inferring Mechanics of Web Censorship Around the World", 2012, <https://www.usenix.org/system/files/conference/foci12/foci12-final1.pdf>.

[Villeneuve-2011]
Villeneuve, N., "Open Access: Chapter 8, Control and Resistance, Attacks on Burmese Opposition Media", 2011, <http://access.opennet.net/wp-content/uploads/2011/12/accesscontested-chapter-08.pdf>.

[VonLohmann-2008]
VonLohmann, F., "FCC Rules Against Comcast for BitTorrent Blocking", 2008, <https://www.eff.org/deeplinks/2008/08/fcc-rules-against-comcast-bit-torrent-blocking>.

[Wagner-2009]
Wagner, B., "Deep Packet Inspection and Internet Censorship: International Convergence on an 'Integrated Technology of Control'", 2009, <http://advocacy.globalvoicesonline.org/wp-content/uploads/2009/06/deeppacketinspectionandinternet-censorship2.pdf>.

[Wagstaff-2013]
Wagstaff, J., "In Malaysia, online election battles take a nasty turn", 2013, <http://www.reuters.com/article/2013/05/04/uk-malaysia-election-online-idUKBRE94309G20130504>.

[Weaver-2009]
Weaver, N., Sommer, R., and V. Paxson, "Detecting Forged TCP Packets", 2009, <http://www.icir.org/vern/papers/reset-injection.ndss09.pdf>.

[Whittaker-2013]
Whittaker, Z., "1,168 keywords Skype uses to censor, monitor its Chinese users", 2013, <http://www.zdnet.com/1168-keywords-skype-uses-to-censor-monitor-its-chinese-users-7000012328/>.

[Wikip-DoS]
Wikipedia, "Denial of Service Attacks", 2016, <https://en.wikipedia.org/w/index.php?title=Denial-of-service_attack&oldid=710558258>.

[Wilde-2012]
Wilde, T., "Knock Knock Knockin' on Bridges Doors", 2012, <https://blog.torproject.org/blog/knock-knock-knockin-bridges-doors>.

[Winter-2012]
Winter, P., "How China is Blocking Tor", 2012, <http://arxiv.org/pdf/1204.0447v1.pdf>.

[Zhu-2011]
Zhu, T., "An Analysis of Chinese Search Engine Filtering", 2011, <http://arxiv.org/ftp/arxiv/papers/1107/1107.3794.pdf>.

[Zmijewki-2014]
Zmijewki, E., "Turkish Internet Censorship Takes a New Turn", 2014, <http://www.renesys.com/2014/03/turkish-internet-censorship/>.

Authors' Addresses

Joseph Lorenzo Hall
CDT

Email: joe@cdt.org

Michael D. Aaron
CU Boulder

Email: michael.aaron@colorado.edu

Stan Adams
CDT

Email: sadams@cdt.org

Ben Jones
Princeton

Email: bj6@cs.princeton.edu

Nick Feamster
Princeton

Email: feamster@cs.princeton.edu

Network Working Group
Internet-Draft
Intended status: Informational
Expires: 15 July 2024

I. R. Learmonth
HamBSD
G. Grover
Centre for Internet and Society
M. Knodel
Center for Democracy and Technology
12 January 2024

Guidelines for Performing Safe Measurement on the Internet
draft-irtf-pearg-safe-internet-measurement-09

Abstract

Internet measurement is important to researchers from industry, academia and civil society. While measurement of the internet can give insight into the functioning and usage of the internet, it can present risks to user privacy and safety. This document describes briefly those risks and proposes guidelines for ensuring that internet measurements can be carried out safely, with examples.

Note

This document is a draft. It is not an IETF product. It does not propose a standard. Comments are solicited and should be addressed to the research group's mailing list at pearg@irtf.org and/or the author(s).

The sources for this draft are at:

https://github.com/IRTF-PEARG/draft-safe-internet-measurement

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 15 July 2024.

Copyright Notice

Copyright (c) 2024 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (https://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document.

Table of Contents

1. Introduction

Measurement of the internet provides important insights and is a growing area of research. Similarly, the internet plays a role in enhancing research methods of different kinds.

Performing research using the internet, as opposed to an isolated testbed or simulation platform, means that experiments co-exist in a space with other services and end users. Furthermore privacy

considerations are of particular importance in internet measurement research that depends on collaboration and data sharing models between industry and academia[caida].

This document outlines guidelines for academic, industry and civil society researchers who might use the internet as part of scientific experimentation to mitigate risks to the safety of users.

1.1. Scope of this document

These are guidelines for how to measure the internet safely. When performing research on a platform shared with live traffic from other users, that research is considered safe if and only if other users are protected from or unlikely to experience danger, risk, or injury arising due to the research, now or in the future.

Following the guidelines contained within this document is not a substitute for institutional ethics review processes, although these guidelines could help to inform that process. It is particularly important for the growing area of research that includes internet measurement to better equip review boards to evaluate internet measurement methods [SIGCOMM], and we hope that this document is part of that larger effort.

Similarly, these guidelines are not legal advice and local laws must also be considered before starting any experiment that could have adverse impacts on user safety.

The scope of this document is restricted to guidelines that mitigate exposure to risks to user safety when measuring properties of the internet: the network, its constituent hosts and links, or user traffic.

1.2. Terminology

Threat model: A threat is a potential for a security violation, which exists when there is a circumstance, capability, action, or event that could breach security and cause harm [RFC4949].

User: For the purpose of this document, an internet user is an individual or organisation whose data is used in communications over the internet, most broadly, and those who use the internet to communicate or maintain internet infrastructure.

Active measurement: Active measurements generate or modify traffic.

Passive measurement: Passive measurements involve the observation of existing traffic without active intervention.

On/off-path: A measurement that is on-path happens on the network. Off-path indicates activity in a side-channel, end-point or at other points where the user, their connection, or their data can be accessed.

One-/two-ended: A single-ended measurement is like a probe or a trace, whereas a measurement with two-ended control provides more accuracy but requires the cooperation of both endpoints, which might include the network itself if that is the measurement target.

1.3. User impact from measurement studies

Any conceivable internet measurement study might have an impact on an internet user's safety. The measurement of generated traffic may also lead to insights into other users' traffic indirectly as well. It is always necessary to consider the best approach to mitigate the impact of measurements, and to balance the risks of measurements against the benefits to impacted users.

Some possible ways in which users can be affected as a result of an internet measurement study:

Breach of privacy: User privacy can be violated in the context of data collection. This impact also covers the case of an internet user's data being shared beyond that for which a user had given consent. First-order data that distinguishes a person such as name, as well as second-order data that can be used to track behaviour such as IP address, should be considered[Kenneally]

Inadequate data protection: A scenario where data, either in transit or at rest, lacks sufficient protection from disclosure. Failure to meet user expectations for data protection is a concern, even if it does not result in unauthorized access to the data. This includes cases of improper access control (i.e. people having access to user data who do not need it).

Traffic generation: A scenario where undue traffic is generated to traverse the internet.

Traffic modification: A scenario where users' on-path internet traffic is nonconsensually modified.

Impersonation: A scenario where a user is impersonated during a measurement.

Legal: Users and service providers are bound by a wide range of policies from Terms of Service to rule of law, each according to context and jurisdiction. A measurement study may violate these policies, and the consequences of such a violation may be severe.

Unavailability: Users or other entities may rely on the information or systems that are involved in the research and they may be harmed by unexpected or planned unavailability of that information or systems[Menlo].

System or data corruption: A scenario where generated or modified traffic causes the corruption of a system. This covers cases where a user's data may be lost or corrupted, and cases where a user's access to a system may be affected as a result.

Emotional trauma: A scenario where a measurement of or exposure to content or behaviour in an internet measurement study causes a user emotional or psychological harm.

2. Guidelines

2.1. Attribute

Proactively identify your measurement to others on the network. "This allows any party or organization to understand what an unsolicited probe packet is, what its purpose is, and, most importantly, who to contact."[RFC9511]

Example: For a layer 3 IP packet probe you could mark measurements with a probe description URI as defined in RFC9511.

2.2. Obtain consent

Accountability and transparency are fundamentally related to consent. As per the Menlo Report, "Accountability demands that research methodology, ethical evaluations, data collected, and results generated should be documented and made available responsibly in accordance with balancing risks and benefits."[Menlo] A user is best placed to balance the risks and benefits for themselves therefore consent must be obtained. From most transparent to least, there are a few options for obtaining consent.

2.2.1. Informed consent

Informed consent should be collected from all users that may be placed at risk by an experiment.

For consent to be informed, a reasonable coverage of possible risks must be presented to the users. The considerations in this document can be used to provide a starting point although other risks may be present depending on the nature of the measurements to be performed. In addition, it should be clear from the consent language who the asker is, and what the terms of data observation and/or collection are.

Example: A researcher would like to use volunteer-owned mobile devices to collect information about local internet censorship. Connections will be attempted by the volunteer's device with services and content known or suspected to be subject to censorship orders.

This experiment can carry substantial risk for the user depending on their specific circumstances. Trying to access censored material can be seen as (network) policy infringement or breaking laws. Consequences can range from disciplinary action from their employer to arrest or imprisonment by government authorities. If the experimenter wants to expose volunteers to this kind of risk, users must be fully informed, and voluntarily give consent to run the measurement. Even then, experimenters should seriously consider designing their experiment in another way.

Note that informed consent is notoriously tricky to obtain. Conveying all possible risks of a measurement is often simply impractical, depending upon how technical the user audience is, the context of the consent prompt, what the tool is normally used by users for, etc. In addition, consent can have network effects. For example, asking a user to consent to sharing information about their communication with others can have impacts on users who have not personally consented to the study.

2.2.2. Proxy consent

In cases where it is not practical to collect informed consent from all users of a shared network, it may be possible to obtain proxy consent. Proxy consent may be given by a network operator or employer that would be more familiar with the expectations of users of a network than the researcher.

In some cases, a network operator or employer may have terms of service that specifically allow for giving consent to third parties to perform certain experiments.

Example: Some researchers would like to perform a packet capture to determine the TCP options and their values used by all client devices on a corporate wireless network.

The employer may already have terms of service laid out that allow them to provide proxy consent for this experiment on behalf of the employees, in this case the users of the network. The purpose of the experiment may affect whether or not they are able to provide this consent. Say, performing engineering work on the network may be allowed, whereas academic research may not be already covered.

Example: A research project looks at networked "things", yet users' only interface with the network is through a device that does not provide interaction to the degree that would be sufficient to obtain informed consent at time of use.

However in this case the user can be informed of the use of data for internet measurement research in the device's terms of use and privacy notice, which can be included in a printed, physical manual for the device or accessed at any time via a webpage. These are examples of proxy consent such that the device manufacturer may choose to share data under certain specified conditions, or to conduct their own measurements.

2.2.3. Implied consent

In larger scale measurements, even proxy consent collection may not be practical. In this case, implied consent may be presumed from users for some measurements. Consider that users of a network will have certain expectations of privacy and those expectations may not align with the privacy guarantees offered by the technologies they are using. As a thought experiment, consider how users might respond if asked for their informed consent for the measurements you'd like to perform.

Implied consent should not be considered sufficient for any experiment that may collect sensitive or personally identifying information. If practical, attempt to obtain informed consent or proxy consent from a sample of users to better understand the expectations of other users.

Example: A researcher would like to run a measurement campaign to determine the maximum supported TLS version on popular web servers.

The operator of a web server that is exposed to the internet hosting a popular website would have the expectation that it may be included in surveys that look at supported protocols or extensions but would not expect that attempts be made to degrade the service with large numbers of simultaneous connections.

Example: A researcher would like to perform A/B testing for protocol feature and how it affects web performance. They have created two versions of their software and have instrumented both to report telemetry back. These updates will be pushed to users at random by the software's auto-update framework. The telemetry consists only of performance metrics and does not contain any personally identifying or sensitive information.

As users expect to receive automatic updates, the effect of changing the behaviour of the software is already expected by the user. If users have already been informed that data will be reported back to the developers of the software, then again the addition of new metrics would be expected. Note that the reduced impact of A/B testing should not be used be an excuse to push updates that might compromise user expectations around security and privacy.

In the event that something does go wrong with the update, it should be easy for users to discover that they have been part of an experiment and roll back the change, allowing for explicit refusal of consent to override the presumed implied consent.

2.3. Share responsibly

Further to use of measurement data, data is often shared with other researchers. Measurement data sharing comes with its own set of expectations and responsibilities of the provider. Likewise there are responsibilities that come with the use of others measurement data. One obvious expectation is around end-user consent (see "Implied consent" above). Allman and Paxson [Allman] provide "a set of guidelines that aim to aid the process of sharing measurement data... [in] a framework under which providers and users can better attain a mutual understanding about how to treat particular datasets."

Their guidance since 2007 has been for data providers to:

* explicitly indications of the terms of a datasets acceptable use

* convey what interactions they desire or will accommodate.

Their guidance for researchers is to:

* be thoughtful in the reporting of potentially sensitive information gleaned from providers data.

* comply with the indications and interactions of the data providers.

Example: Researchers have obtained network measurement data from more than one provider for purposes of conducting analysis of protocol use on both. Where privacy paritioning techniques are used, the researchers' findings may inadvertently collude to uncover private information about users. Once realised, researchers should mitigate this privacy risk to end users as well as disclosing this result to the data providers themselves.

2.4. Isolate risk with a dedicated testbed

Wherever possible, use a testbed. An isolated network means that there are no other users sharing the infrastructure you are using for your experiments.

When measuring performance, competing traffic can have negative effects on the performance of your test traffic and so the testbed approach can also produce more accurate and repeatable results than experiments using the public internet.

Example: WAN link conditions can be emulated through artificial delays and/or packet loss using a tool like [netem]. Competing traffic can also be emulated using traffic generators.

2.5. Be respectful of others' infrastructure

If your experiment is designed to trigger a response from infrastructure that is not your own, consider what the negative consequences of that may be. At the very least your experiment will consume bandwidth that may have to be paid for.

In more extreme circumstances, you could cause traffic to be generated that causes legal trouble for the owner of that infrastructure. The internet is a global network that crosses many legal jurisdictions and so what may be legal for one is not necessarily legal for another.

If you are sending a lot of traffic quickly, or otherwise generally deviating from typical client behaviour, a network may identify this as an attack which means that you will not be collecting results that are representative of what a typical client would see.

One possible way to mitigate this risk is transparency, i.e. mark measurement-related data or activity as such. For example, the popular internet measurement tool ZMap hardcodes its packets to have IP ID 54321 in order to allow identification [ZMap].

2.6. Maintain a "Do Not Scan" list

When performing active measurements on a shared network, maintain a list of hosts that you will never scan regardless of whether they appear in your target lists. When developing tools for performing active measurement, or traffic generation for use in a larger measurement system, ensure that the tool will support the use of a "Do Not Scan" list.

If complaints are made that request you do not generate traffic towards a host or network, you must add that host or network to your "Do Not Scan" list, even if no explanation is given or the request is automated.

You may ask the requester for their reasoning if it would be useful to your experiment. This can also be an opportunity to explain your research and offer to share any results that may be of interest. If you plan to share the reasoning when publishing your measurement results, e.g. in an academic paper, you must seek consent for this from the requester.

Be aware that in publishing your measurement results, it may be possible to infer your "Do Not Scan" list from those results. For example, if you measured a well-known list of popular websites then it would be possible to correlate the results with that list to determine which are missing. This inference might leak the fact that those websites specifically requested to not be scanned.

2.7. Minimize data

When collecting, using, disclosing, and storing data from a measurement, use only the minimal data necessary to perform a task. Reducing the amount of data reduces the amount of data that can be misused or leaked.

When deciding on the data to collect, assume that any data collected might be disclosed. There are many ways that this could happen, through operational security mistakes or compulsion by a judicial system.

When directly instrumenting a protocol to provide metrics to a passive observer, see section 6.1 of RFC6973[RFC6973] for the data minimization considerations enumerated below that are specific to the use case.

2.7.1. Discard data

Discard data that is not required to perform the task.

When performing active measurements, be sure to only capture traffic that you have generated. Traffic may be identified by IP ranges or by some token that is unlikely to be used by other users.

Again, this can help to improve the accuracy and repeatability of your experiment. For performance benchmarking, [RFC2544] requires that any frames received that were not part of the test traffic are discarded and not counted in the results.

2.7.2. Mask data

Mask data that is not required to perform the task. This technique is particularly useful for content of traffic to indicate that either a particular class of content existed or did not exist, or the length of the content, but not recording the content itself. The content can be replaced with tokens or encrypted.

It is important to note that masking data does not necessarily anonymize it [SurveyNetworkTrafficAnonymisationTech].

2.7.3. Aggregate data

When collecting data, consider if the granularity can be limited by using bins or adding noise. Differential privacy techniques [DifferentialPrivacy] can help with this.

Example: [Tor.2017-04-001] presents a case-study on the in-memory statistics in the software used by the Tor network.

2.8. Reduce accuracy

There are various techniques that can be used to reduce the accuracy of the collected data and make it less identifying.

The use of binning to group numbers of more-or-less continuous values, coarse categorization in modeling, reduction in concentrations of IP address by geography (geoip) or other first- or second-order identifiers, the introduction of noise and all privacy-preserving measurement techniques that allow researchers to safely conduct internet measurement experiments without risking harm to real users[Janson].

2.9. Analyze risk

The benefits of internet measurement should outweigh the risks. Consider auxiliary data (e.g. third-party data sets) when assessing the risks. Consider that while a privacy risk may not be immediately apparent or realisable, in the future increased computing power may then make something possible.

Example: A research project releases encrypted payloads as a method for minimising exposure of sensitive user data. However the encryption could be trivially broken in the future with typical increases in computing power.

3. Security Considerations

This document as a whole addresses user safety considerations for internet measurement studies, and thus discusses security considerations extensively throughout regarding collection and storage of user data.

4. IANA Considerations

This document has no actions for IANA.

5. Acknowledgements

Many of these considerations are based on those from the [TorSafetyBoard] adapted and generalised to be applied to internet research.

Other considerations are taken from the Menlo Report [Menlo] and its companion document [MenloReportCompanion].

Comments of several people on the mailing list was helpful, especially Marwan Fayed and Jeroen van der Ham.

6. Informative References

[netem] Stephen, H., "Network emulation with NetEm", April 2005.

[RFC2544] Bradner, S. and J. McQuaid, "Benchmarking Methodology for Network Interconnect Devices", RFC 2544, DOI 10.17487/RFC2544, March 1999, <https://www.rfc-editor.org/info/rfc2544>.

[TorSafetyBoard] Tor Project, "Tor Research Safety Board", <https://research.torproject.org/safetyboard/>.

[RFC4949] Shirey, R., "Internet Security Glossary, Version 2", August 2007, <https://www.rfc-editor.org/info/rfc4949>.

[Tor.2017-04-001] Herm, K., "Privacy analysis of Tor's in-memory statistics", Tor Tech Report 2017-04-001, April 2017, <https://research.torproject.org/techreports/privacy-in-memory-2017-04-28.pdf>.

[Menlo] Dittrich, D. and E. Kenneally, "The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research", August 2012, <https://www.dhs.gov/sites/default/files/publications/CSD-MenloPrinciplesCORE-20120803_1.pdf>.

[MenloReportCompanion] Bailey, M., Dittrich, D., and E. Kenneally, "Applying Ethical Principles to Information and Communication Technology Research", October 2013, <https://www.impactcybertrust.org/link_docs/Menlo-Report-Companion.pdf>.

[DifferentialPrivacy] Dwork, C., McSherry, F., Nissim, K., and A. Smith, "Calibrating Noise to Sensitivity in Private Data Analysis", 2006, <https://link.springer.com/chapter/10.1007/11681878_14>.

[SurveyNetworkTrafficAnonymisationTech] Van Dijkhuizen, N. and J. Van Der Ham, "A Survey of Network Traffic Anonymisation Techniques and Implementations", May 2018, <https://dl.acm.org/doi/10.1145/3182660>.

[ZMap] University of Michigan, "ZMap Source Code - packet.c", <https://github.com/zmap/zmap/blob/main/src/probe_modules/packet.c>.

[RFC6973] Cooper, A., Tschofenig, H., Aboba, B., Peterson, J., Morris, J., Hansen, M., and R. Smith, "Privacy Considerations for Internet Protocols", RFC 6973, July 2013, <https://www.rfc-editor.org/info/rfc6937>.

[SIGCOMM] Jones, B., Ensafi, R., Feamster, N., Paxson, V., and N. Weaver, "Ethical Concerns for Censorship Measurement", August 2015, <http://conferences.sigcomm.org/sigcomm/2015/pdf/papers/nsethics/p17.pdf>.

[RFC9511] Vyncke, É., Donnet, B., and J. Iurman, "Attribution of Internet Probes", November 2023, <https://www.rfc-editor.org/info/rfc9511>.

[Allman] Allman, M. and V. Paxson, "Issues and Etiquette Concerning Use of Shared Measurement Data", October 2007, <https://conferences.sigcomm.org/imc/2007/papers/imc80.pdf>.

[caida] CAIDA, "Promotion of Data Sharing", January 2010, <https://www.caida.org/catalog/datasets/sharing>.

[Kenneally] Kenneally, E. and K. Claffy, "Dialing privacy and utility: a proposed data-sharing framework to advance Internet research", 2010, <https://www.caida.org/catalog/papers/2010_dialing_privacy_utility/dialing_privacy_utility.pdf>.

[Janson] Janson, R., Traudt, M., and N. Hopper, "Privacy-Preserving Dynamic Learning of Tor Network Traffic", 2010, <https://dl.acm.org/doi/pdf/10.1145/3243734.3243815>.

Authors' Addresses

Iain R. Learmonth
HamBSD
Email: irl@hambsd.org

Gurshabad Grover
Centre for Internet and Society
Email: gurshabad@cis-india.org

Mallory Knodel
Center for Democracy and Technology
Email: mknodel@cdt.org

Network Working Group
Internet-Draft
Intended status: Experimental
Expires: January 14, 2021

S. Rao
S. Nagaraj
Grab
S. Sahib
R. Guest
Salesforce

July 13, 2020

Personal Information Tagging for Logs
draft-rao-pitfol-02

Abstract

Software systems typically generate log messages in the course of their operation. These log messages (or 'logs') record events as they happen, thus providing a trail that can be used to understand the state of the system and help with troubleshooting issues. Given that logs try to capture state that is useful for monitoring and debugging, they can contain information that can be used to identify users. Personal data identification and anonymization in logs is crucial to ensure that no personal data is being inadvertently logged and retained which would make the logging system run afoul of laws around storing private information. This document focuses on exploring mechanisms that can be used by a generating or intermediary logging service to specify personal or sensitive data in log message(s), thus allowing a downstream logging server to potentially enforce any redaction or transformation.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any

time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on January 14, 2021.

Copyright Notice

Copyright (c) 2020 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (https://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction

Logs capture the state of a software system in operation, thus providing observability. However, because of the amount of state they capture, they can often contain sensitive user information [link: twitter storing passwords]. Personal data identification and redaction is crucial to make sure that a logging application is not storing and potentially leaking users' private information. There are known precedents that help discover and extract sensitive data, for example, we can define a regular expression or lookup rules that will match a person's name, credit card number, email address and so

on. Besides, there are data dictionary based training models that can analyze logs and predict presence of sensitive data and subsequently redact it. This document proposes an approach and framework for creating logs with personal information tagged, thus marking a step towards privacy aware logging. Once personal information is identified in a log, it has to be appropriately tagged at source. Personal data tagging is especially important in cases where log data is flowing in from disparate sources. In cases where tagging at source is not possible (e.g. log data generated by a legacy application, IoT device, Web server or a Firewall), a centralized logging server can be tasked with making sure the log data is tagged before passing on downstream. Once the logs are tagged, the logging application can use anonymization techniques to redact the fields appropriately. While the proposal described here can be applied to any data deemed sensitive in a log, however this document specifically discusses and illustrates tagging of personal information in logs.

2. Terminology

Personal data: RFC 6973 [RFC6973] defines personal data as "any information relating to an individual who can be identified, directly or indirectly." This typically includes information such as IP addresses, username, email address, financial data, passwords and so on. However, the definition of personal data varies heavily by what other information is available, the jurisdiction of operation and other such factors. Hence, this document does not focus on prescriptively listing what log fields contain personal data but rather on what a tagging mechanism would look like once a logging application has determined which fields it considers to hold personal data.

Structured logging: Most applications generate logs in a unidimensional format that twine together logic status and input data. This makes log output largely free flowing and unstructured without specific delimiters making it hard to segregate personal information from other text in the log. Structured logging refers to a formal arrangement of logs with specific identifiers of personal information and semantic information to enable easy parsing and identification of specific information in the log.

Privacy Sensitivity Level: Sensitivity level defines the degree of sensitivity of a data in log template or schema. Level can be enumerated on a scale 1 to 5 and defined as follows: 1 - Low risk for leaking private information and 5 - Very high risk for leaking private information>

3. Motivation and Use Cases

Most systems like network devices, web servers and application services record information about user activity, transactions, network flows, etc., as log data. Logs are incredibly useful for various purposes such as security monitoring, application debugging, investigations and operational maintenance. In addition, there are use cases of organizations exporting or sharing logs with third party log analyzers for purposes of security incident response, monitoring, business analytics, where logs can be a valuable source of information. In such cases, there are concerns about potential exposure of personal data to unintended systems or recipients.

4. Challenges with Existing Approaches

While methods of detecting personal identifiable information are continuously evolving, most approaches are around use of regular expressions, data or dataset based training models, pattern recognition, checksum matching, building custom logic.

Inconsistent Representation: When applications, services or devices, log personal information, there is no consistency in the representation of the information. For example the name of a user is often logged as either "fullname" (e.g. John Doe) or with "firstname" (John) and "lastname" (Doe).

Context: In most cases, what data is considered personal and sensitive is subjective, provisional and contextual to the data source or the application processing the data, which makes it hard to use automated techniques to identify personal data. Even for a specific domain, it’s controversial whether it is possible to definitively say that a piece of data is NOT identifying.

Disparate Types of Personal Data: There are many disparate types of personal data and often require a multitude approaches for detection.

Lack of standards: There are no standards that govern formats of sensitive data making automation difficult for most common use cases.

Detection Accuracy: Most of the current PII detections tools employ regular expression based techniques or other pattern recognition techniques to identify the PII data. Due to the very nature of logs, most of the current implementations let administrators to add redaction policies based on ’likelihood’ of detection probability categorized as low, medium or high. Defining a low detection scheme causes high false positives and a high detection scheme would cause PII leakage, thereby making a trade off inevitable to organizations.

5. Proposed Model

This section describes a reference model to enable tagging of personal information at source and extends it to include an approach of role or policy based redaction based on personal information annotated at source. The figure below illustrates the proposed model.

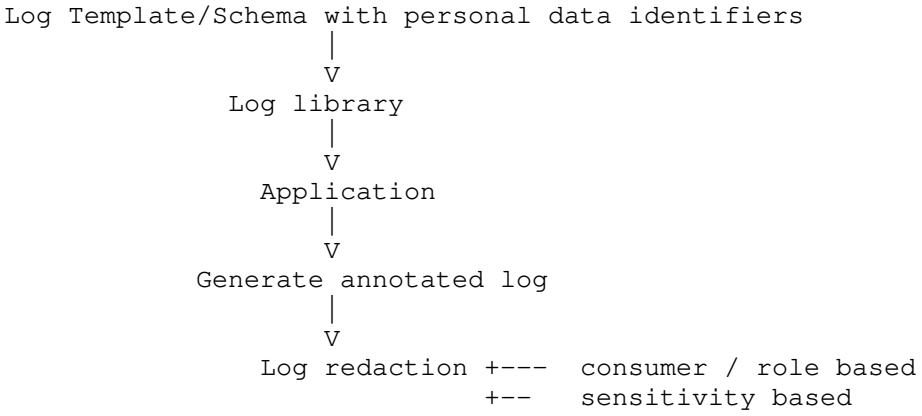

Figure 1: Flow

5.1. Defining the log privacy schema

We propose using structured logging where a log schema or a template defines standardized identifiers for every personal information and each log field is associated with a sensitivity level customized to a use case or log intent.

Note that this is not to be confused with a log severity level (WARN, INFO...) - those are typically defined "dynamically" by the developer while defining the severity of a certain scenario. A privacy sensitivity level is defined statically and is part of a log schema, associated with the log name and data type.

Name	Abstract Data Type	Description	Sensitivity [1-High 5-Normal]
nationalIdentity	String	National IDs issued by sovereign governments. Eg., SSN	1
drivingLicense	String	Driving License number	1
taxIdentity	String	Tax identification numbers	1
credtCardNumber	String	Credit cards	1
bankAccount	String	Bank account number	1
dateOfBirth	Date	Date of Birth	2
personName	String	Person name	1
emailAddress	String	Email	2
phoneNumber	Number	Phone	1
zipCode	Integer	Zip codes	5
ipAddress	ipv4Address	IPv4 or IPv6 Address	4
dateTimeSeconds	dateTimeSeconds	seconds	5
age	Integer	Age	2
ethnicGroup	String	Ethnic group	1
genderIdentity	String	Gender identity	1
macAddress	macAddress	MAC Address	4

Personal Information Identifiers Registry

If an organization already uses structured logging with a log schema, then a privacy sensitivity level can be an additional attribute for the schema.

The privacy sensitivity level for log types is intended to be defined by a centralized effort around privacy preservation in logs. In other words, this mapping might be done by an organization's privacy team (which can include lawyers, engineers and privacy professionals). The intention is that all logs generated by an org should conform to this structured format, which would ease downstream processing of logs for access control and removal of sensitive information.

If the log is being generated by a web server, then two approaches can be taken:

1. Modify log-format for the service: identify the log data type of each piece of log data generated, and tag in generation (examples provided in later section)

2. Add automated tagging in a centralized log aggregator: collect all the logs generated by different services and apply the annotation using the log schema at the aggregator

5.2. Typical Workflow

1. The log privacy schema can be parsed into a structured logging library, that is used by individual developer teams. The intention is for developers to not log arbitrary data i.e. they are asked to identify what is the data type of the state they want to preserve.

2. Any addition to the log schema would have to go through review of the privacy team that came up with the log schema.

3. Once a log is generated, tagged and stored, various kinds of access control techniques can be applied to who can access the logs.

5.3. Log Processing and Access Control

1. Consumer Role Based Access

    A. Once the log is tagged, access to it can be based on a consumer's role and privilege level.

    B. A consumeer role based policy can define what level of sensitivity they can access.

2. Case-based access

    A. If there is a genuine case for which access to sensitive information is needed and granted by the legal department, a cryptographically-signed token (e.g.JWT) can be generated that will allow access to a developer/user to logs of an increased log level. This access can be temporal in nature i.e. the token will only be valid for a certain amount of time.

    B. A transaction ID can also be propagated automatically throughout the request processing, to correlate different

logs related to a single request. Note that the notion of a "request" can vary based on what the application is doing. The idea is to have a single unifying ID to tie a particular action. If this is done, then the temporary token can be restricted to a particular request ID.

3. Redaction Techniques

A. Given that the log is tagged, an organization might choose to redact the more sensitive logs i.e. ones above a certain sensitivity level, ones of a certain log type.

B. More sophisticated approaches can be developed i.e. completely redact log types username and email, but obfuscate IP address so that a rough location can be garnered from the log record. In this way, techniques such as differential privacy can be used in tandem to have privacy guarantees for logs while still providing usefulness to developers.

6. Examples

An example based on RFC 3164 Log format

Normal Log Ouput

```
<120> Nov 16 16:00:00 10.0.1.11 ABCDEFG: [AF@0 event="AF-Authority
failure" violation="A-Not authorized to object" actual_type="AF-A"
jrn_seq="1001363" timestamp="20120418163258988000"
job_name="QPADEV000B" user_name="XYZZY" job_number="256937"
err_user="TESTFORAF" ip_addr="10.0.1.21" port="55875"
action="Undefined(x00)" val_job="QPADEV000B" val_user="XYZZY"
val_jobno="256937" object="TEST" object_library="CUS9242"
object_type="*FILE" pgm_name="" pgm_libr="" workstation=""]
```

Log Output with Personal Information Tagging

```
<120> Apr 18 16:32:58 10.0.1.11 QAUDJRN: [AF@0 event="AF-Authority
failure" violation="A-Not authorized to object" actual_type="AF-A"
jrn_seq="1001363" timestamp="20120418163258988000"
job_name="QPADEV000B" {personName="XYZZY" pii_sensitivity_level=1}
job_number="256937" {emailAddress="xyz@foo.com"
pii_sensitivity_level=2] [ip_addr="10.0.1.21"
pii_sensitivity_level=4] port="55875" action="Undefined(x00)"
val_job="QPADEV000B" val_jobno="256937" object="TEST"
object_library="CUS9242" object_type="*FILE" pgm_name="" pgm_libr=""
workstation=""]
```

7. IANA Considerations

IANA can consider defining a new central respository for Personal Information name and identifier registries to used in logging personal information. The personal identifier registry would enumerate namee and identifiers as described in Section 5.1.

8. Security Considerations

It is anticipated that developers will want additional log data types for capturing application logic, and might abuse an existing log type instead of going through the process of adding a new one. In such a case, the log would be incorrectly tagged. This can be mitigated by having stronger typing for the log data types i.e. restricting address to a certain string length instead of storing arbitrary length.

Encouraging developers to think carefully about what kind of data they’re logging is a good practice and will lead to fewer incidents of private data being inadvertently logged. An organization might choose to have an unstructured log type for letting developers log data that truly do not fit anywhere else. This is still better than not having structured privacy-aware logging, because the potential privacy leakage is isolated to one particular field and its use can be monitored.

Having a mapping from log data type to privacy sensitivity will need continuous effort by a privacy team, which might be expensive for an organization.

Log data is often collated, propagated, transformed, loaded into different formats or data models for purposes of analytics, troubleshooting and visualization. In such cases, it is necessary and critical to ensure that personal information tagging and annotations is preserved and forwarded across format transformations.

If the privacy marking or classification changes for a log, for historical logs, the change of privacy classification is applied on subsequent access of the log.

TODO: In case of logs that are not tagged or marked with personal information, an out-of-band mechanism to communicate log template or schema with personal data identifiers can be considered. Such a mechansim can also be used to notify changes to privacy tagging or classification.

9. Acknowledgements

The authors would like to thank everyone who provided helpful comments at the mic at IETF 106 during the PEARG session. Thanks also to Joe Salowey for thoughts on aspects of log transformations, change of privacy classifications, models for privacy marking.

10. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC3164] Lonvick, C., "The BSD Syslog Protocol", RFC 3164, DOI 10.17487/RFC3164, August 2001, <https://www.rfc-editor.org/info/rfc3164>.

[RFC6973] Cooper, A., Tschofenig, H., Aboba, B., Peterson, J., Morris, J., Hansen, M., and R. Smith, "Privacy Considerations for Internet Protocols", RFC 6973, DOI 10.17487/RFC6973, July 2013, <https://www.rfc-editor.org/info/rfc6973>.

Authors' Addresses

Sandeep Rao
Grab
Bangalore
India

Email: sandeeprao.ietf@gmail.com

Santhosh C N
Grab

Email: santoshcn1@gmail.com

Shivan Sahib
Salesforce

Email: shivankaulsahib@gmail.com

Ryan Guest
Salesforce

Email: rguest@salesforce.com