The Great Firewall of China

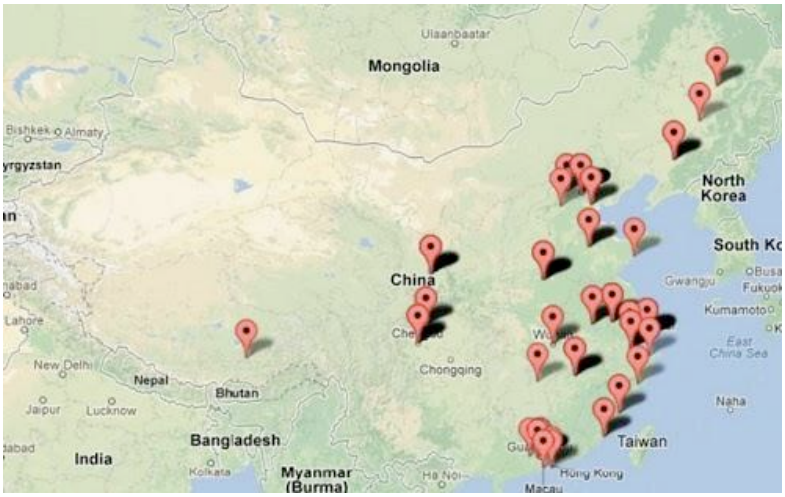
Ruiwei Bu
CSC 540

1. Definition

The Great Firewall of China, or simplified as Great Firewall (GFW), is part of China's "Golden Shield" Project. The project mainly focusing on Internet security, control and censorship. The name comes from a paper named *The Great Firewall of China* by Charles R. Smith in 2002. However, the original paper is lost in the Internet and not accessible anymore. Construction of the system was started in 1998 and China is continuing enhancing the system till now. Major functionalities of the firewall includes DNS pollution and injection, IP and port blocking, TCP reset, interfere secure connection, proxy blocking, IPv6 censorship and Email blocking. GFW is well-known by blocking famous foreign websites including Facebook, Twitter and Google in main land China. It has became a invisible cyber wall blocking the Chinese people from the outside world.

2. System Design Overview

The GFW is consist of two major parts - Hardware and Software. The hardware part is mainly based on CISCO's internet devices and forms a intrusion detection system[1]. These hardwares are deployed on the international gateway and local Internet centers of China's major ISPs (Internet Service Provider, smaller ISPs are sharing the same gateway with major ISPs). Till now, no one actually knows the physical position of GFW's hardware devices due to national security. However, there were projects trying to figure out the number and the position of the devices, the most recent one is called mongol.py, and its based on the theory proposed in a paper by the Department of Computer Science and Engineering in University of Michigan[2].

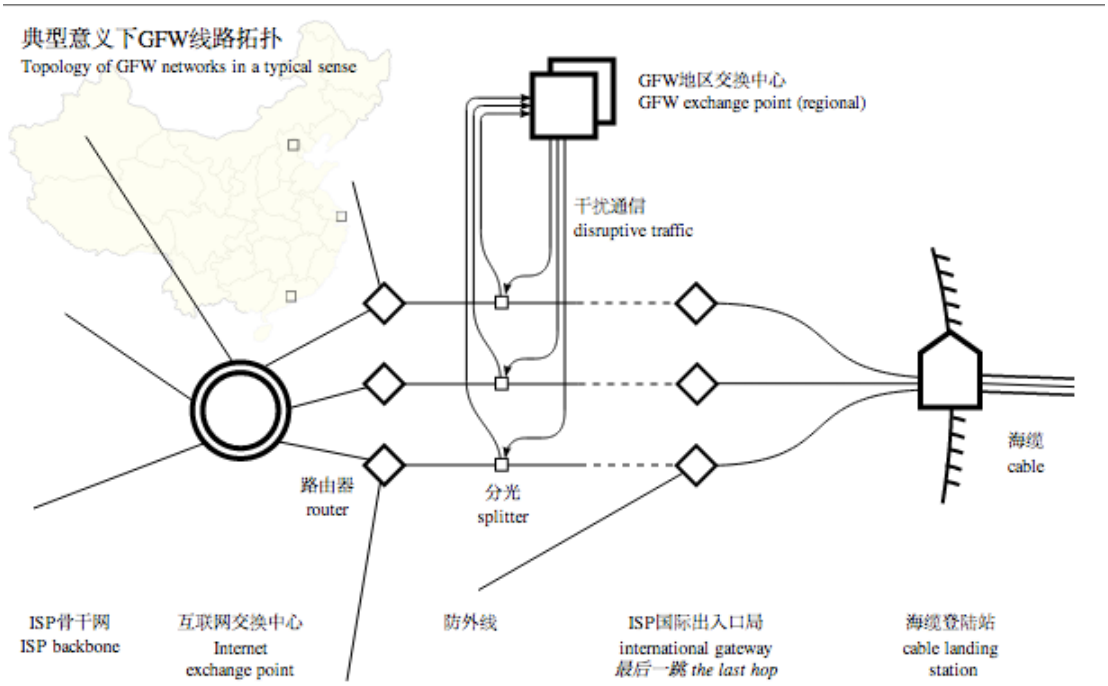


Some hits reported by mongol.py, from http://m.letscorp.net/archives/42652

[1] Overview of the Public Security Sector, Cisco Systems, Inc.

[2] Xueyang Xu, Z. Morley Map, and J. Alex Halderman, Internet Censorship in China: Where Does the Filtering Occur? Department of Computer Science and Engineering, University of Michigan.

The softwares for the GFW mainly comes from some labs in China's top universities and also network related companies. Students worked in the lab may not even know about that they were working for the GFW since on the surface, all the projects are just related to Internet security and censorship. And since the hardware is deployed on the ISPs' international gateways, major network companies in mainland China are cooperating with the government and ISPs in the content censorship. And for cross-country companies such as Microsoft and Cisco, there are special VPNs inside the company for their internal use to bypass the censorship and communication with the headquarters. Most significantly, due to the advances in researches, the softwares of the GFW is getting stronger everyday. At the beginning, they were only able to block websites with keywords. And now, the GFW even has the ability to censor the connection and network flow patterns in popular secured connections such as VPN and SSH tunnels, and thus then block the connections when it thinks its used for proxy[3]. Even more, the GFW is able to know and block the bridges used in the Tor project, including obfsproxy2. Though its not able to block obfsproxy3 currently[4]. Actually, the GFW maybe is the biggest and best place for researches related to internet connections and network flow analysis. Many researchers are using it as part of their research in China's top universities. And usually they can only access the GFW devices deployed in their region, and that is part the reason why sometimes the behavior of the GFW is different in different parts of China.



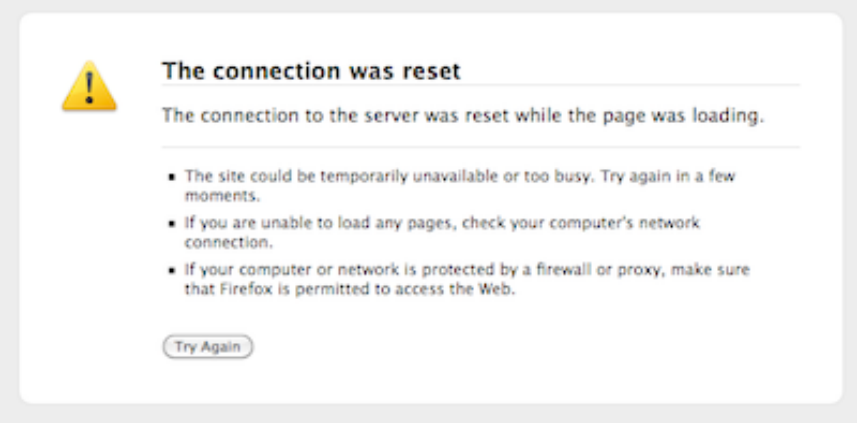
Topology of GFW network in a typical sense
from https://news.ycombinator.com/item?id=4931595

[3] There were many reports and discussions related to VPN and SSH tunnels being blocked by GFW after a recent update in Spring 2013, such as Hacker News: https://news.ycombinator.com/item?id=4931595

[4] https://blog.torproject.org/category/tags/gfw, https://trac.torproject.org/projects/tor/ticket/8591, https://trac.torproject.org/projects/tor/ticket/8591

3. What can the GFW do?

The GFW maybe is the biggest internet censorship network ever created in human's history. It has the power to control and censor all network connections in mainland China based on but not limited to keyword analysis, connection and network flow pattern analysis, special target's IP addresses and URLs and communication datas. The GFW not only listen to network connections goes out of the country, but also internal connections. It's able to censor all of the Chinese network traffic, including but not limited to IM messages, E-mails and website contents, block certain webpages and websites and record all users' activities on the Internet at least for years. A typical example is, when a user in mainland China trying to search sometime not allowed in Google, the user's Internet connection will be blocked for a few minutes. His/Her browser will tell report that it's not able to reach Google's server (Connection Reset), but actually its the GFW have cutdown the connections between the user and Google. The abilities of the GFW includes IP blocking, URL filtering, DNS filtering/polluting, packet filtering, content filtering, SSL certificate faking, Tor node faking, network flow analyzing and even completely block all outside connections and make the Internet of mainland China independent from the outside.



Connection reset when trying to access the blocked websites.

In the Spring of 2013, a fews days before China's Spring Festival, the GFW "accidentally" blocked Github, which is a famous web-based open-source project hosting and Git reversion control system, for a few day. The attack was a famous man in the middle attack, and the GFW replaced Github's SSL certificate with a self-signed certificate[5]. The reason was unclear. Some people believe its because some people are using Github pages as a place to discuss political issues, and caused a huge argument regarding whether we should put political related things on Github or not.

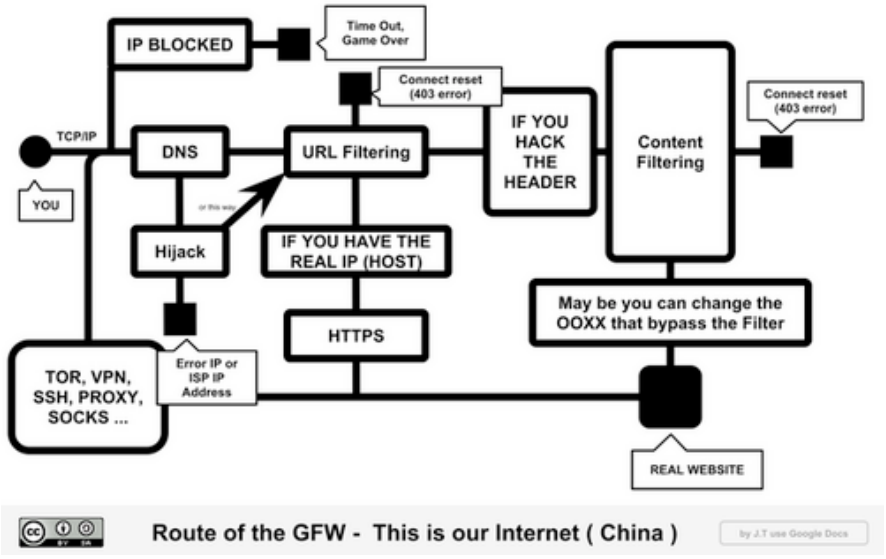
Despite the negative abilities, the GFW might also has some positive abilities, such as protect China's network from cyber attacks. Actually the original attention of the GFW is to protect

[5] Github SLL replaced by self-signed certificate in China. https://news.ycombinator.com/item?id=5124784

China from potential network attacks. But now its mostly used as a censorship tool for the governors.

4. Route of the GFW

Typically the GFW is separated into different connection layers. First, when a user tries to connect to a URL, the GFW checks whether the user's IP is blocked (in blacklist) or not. Then the TCP/IP packets flow to the DNS server. The DNS server maybe polluted by the GFW and returns a error IP address or a webpage says the website is not accessible before the DNS server returns the real IP address. Users can bypass DNS pollutions by manually setting the DNS server of the Internet connection to public DNS servers such as Google DNS and OpenDNS instead of using ISP's DNS servers, or setting up a local DNS server to resolve the URLs locally. Also in the DNS resolution step, the URL maybe filtered by the GFW and return a RST packet[6]. If the user manages to get the real IP address of the website and accessing it with HTTPS connection, he/she maybe able to get to the real website. If the connection is not protected by HTTPS, then the GFW is able to do filter the website's content. If it found anything not permitted, it will also return a RST packet.



Route of the GFW
from http://thenextweb.com/asia/2010/01/05/route-gfw-flow-chart/

5. Target of the GFW

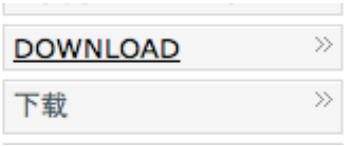
Anything existing on the Internet can be the GFW's target. But the GFW is mostly interested in

- UGC (User Generated Content) and their hosts, such as Twitter, Facebook, Blogspot, Youtube, Fc2 and so on. The list can be extremely long and one of the interesting fact is that most of the blocked services have a similar clone in mainland China. Such as Sina Weibo - a hybrid of Twitter and Facebook, Youku and Tudou - video hosting sites similar to Youtube and RenRen -

[6] TCP Packet Control Bits, RST = Connection Reset. http://www.ietf.org/rfc/rfc793.txt

a direct copy of Facebook.

- Information related to Chinese government and politics. Sometimes unrelated things may also become a target because of the keyword filtering. For example, the name of the formal Chinese president is Hu Jintao (胡锦涛). But when a user trying to search carrot (胡萝卜) with Google, the connection may also be blocked because one of the Chinese character is the same.
- Opinion on public affairs that goes against the current government, such as Tibetan issue.
- Cults, such as Falun Gong.
- Things related to national security.
- Random websites, such as Github, SourceForge and Python's official website. Python's official website even has a special link for Chinese users because the block.



Special link for Chinese users. The Chinese word here means "Download".

Moreover, such huge network can also be used to apply network attacks. For example, DNS injection used in the GFW can cause collateral damage to the root DNS servers[7].

5. People behind the GFW

It's never clear who is actually guiding the deployment and development of the GFW. The public can only know about some of the researchers through the papers published[8], such as Binxing Fnag, who is the principle of BUPT (Beijing University of Post and Telecommunications), a key national university distinguished by the teaching and researching in the field of cable and wireless communications. Fang is recognized by the father of China's Great Firewall because of his research in the cyber security field. And many of his papers are related to the behavior of the GFW. For example, one of the paper signed with his name described the progress aiming at challenges in reality such as real-time classification in backbone network, encrypted traffic classification, fine-grained classification, constantly changing protocols classification and so on[9]. And this paper is thought related to the block of OpenVPN and similar services in China.

---

[7] Anonymous: The collateral damage of internet censorship by DNS injection. CCR July 2012.

[8] Anonymous: Finding the contributors to Great Firewall by their papers

[9] Xiong Gang, Meng Jiao, Cao Zi-gang, Wang Yong, Guo Li, Fang Binxing, Research Progress and Prospects of Network Traffic Classification. Journal of Integration Technology, Vol 1, May, 2012.

In 2011, when Fang was giving a lecture in Wuhan University, which is one of the major universities in China, he was hit by a shoe from the protesters. It's still unknown who did this but the event caused a huge number of responses among major social networks. Although governors quickly blocked all related news and discussions in mainland China and claims it was a crime, they cannot block people outside of the wall.

In the Spring of 2013, when Github was blocked by the GFW, furious programmers posted some names they believed that's related to the GFW as a Gist on Github[10]. And the gist caused a long argument in whether people should put such things on Github or not. Some people is skeptical that these kinds of information will cause long-term block of Github in mainland China while other people think we should not censor ourselves.

6. Ways to bypass the GFW

There are many methods to bypass the GFW in mainland China, and all of them are also ways for people to hide their identity when browsing the Internet. These methods are be separated into four major categories, including *host modification*, *proxy* and *VPN*. Non of the methods is guaranteed to work permanently as the GFW keeps upgrading, and the service providers may not reliable.

1. Host Modification

   This is the simplest method to access blocked websites. Each operating system provides a simple plain text file called *hosts file* and its used to map hostnames to IP addresses. In general IP (Internet Protocol) implementations, before the operating system actually send DNS packets to DNS servers, it will first look at the file to see if the url is in the *hosts file*. If the there is a mapping rule in the file, the operating system will try to use the IP address directly unless otherwise configured. The location of the file varies by operating system[11] and can be directly modified by system administrators.

Operating System	Location
Microsoft Windows	%SystemRoot%/system32/drivers/etc/hosts
Linux, Unix and POSIX (including OS X)	/etc/hosts

   However, due to the fact that the IP addresses of websites may change over time, this method is not reliable and only works when the connection is secured.

2. Proxy

   The principle of a proxy is instead of connecting to the target server directly, a proxy server is used as an intermediary for requests. So if the client machine is blocked while the proxy

[10] https://gist.github.com/shenzhuxi/4635732

[11] http://support.microsoft.com/kb/972034, http://support.apple.com/kb/TA27291?viewlocale=en_US

server is safe and the client can access the proxy server, it can use the proxy server to access the target server and then get the result from the proxy server. There are four major types of proxy, including *tunnel proxy, forward proxy, reverse proxy* and *open proxy.* Each of them works differently and there are many applications in each of them. Some typical application of the proxy method are:

a. Online Proxies
   Online proxies are open proxies and they are the most common and easy to use method to hide one's identity. Usually they are websites that retrieves the content of target website for the users. Though this is the most unsafe way and there are a variety of methods to detect.

b. Proxy Softwares
   These are softwares that connects to proxy servers or sometimes users around the world to hide the users' identities. Usually they use private protocols when communicating with the proxy servers and its harder to detect. They are similar to private proxies but easier to use. This is another common and easy to use method as the user only needs to run a special software and it will do everything for the user. Usually the softwares are packed with a customized browser for "safe browsing". Though the users can also configure their favorite browser to use a local proxy and achieve the same result.

   Freegate and Wujie are two popular proxy softwares in mainland China. The developer and fundings of these softwares are still unknown. Usually they are delivered by E-mails through mail-lists or automatic responses. And these softwares needs constant upgrades due to the fact that the GFW is able to block the proxy servers used by them. There are some rumors that they are funded by west governments and used as a tool to spread political news about the Chinese government (actually the start-up page of the customized browser is a famous rumor site).

   Another popular software is Tor, which is famous all over the world. Its created for anonymous Internet accessing and its similar to a P2P network. The software redirects internet traffic to a worldwide volunteer network of servers (or users?) and the original data is encrypted multiple times during the transition. So its also called "Onion Network" or "Onion Routing" due to the layered nature. It also builds a pseudo top-level domain ".onion". Websites with this domain can only be accessed within the Tor network, and its often used for high-anomaly communities, such as drag transactions and illegal sexual contents. Till now, the GFW is able to probe and block obfs2 bridges used in the Tor network, but still not able to block obfs3 bridges[12].

c. Private Proxies (Tunneling)
   Advanced users can deploy their own proxy servers and use private protocols for the

[12] GFW actively probes obfs2 bridges. https://trac.torproject.org/projects/tor/ticket/8591

communications to bypass the GFW. A common way to do this is to purchase a VPS (Virtual Private Server) from reliable providers and play with it. Private proxies are much safer as they are not largely deployed and its totally controlled by the user. But it also requires more advanced skills in networking and server administration.

SSH (Secure Shell) Tunnel is a typical application of the tunneling protocol. It consists of an encrypted tunnel created over a SSH connection. SSH is a another protocol for secure data communication. It can forward data from a local port to another port on the remote machine, so it can be used to bypass firewalls as the connection goes on. For example, a user can configure a SSH tunnel that forwards data from a local port to port 80 (HTTP) or 443 (HTTPS) on an external SSH server (such as a VPS) to bypass GFW's block in the local network. Usually VPSs come with SSH connection enabled or users can install and configure such capability by themselves.

There are also open-source tunnel proxies that can be deployed on the servers, such as *shadowsocks* by clowwindy[13].

Google App Engine can also be used as a proxy server. As Google provides a free monthly quota for each app, it can be used as a free proxy server. One famous implementation of the proxy based Google App Engine is called GoAgent. It's a famous software based on Python for communicating with Google's App Engine servers[14].

However, this method is also not always reliable as the GFW may block the VPS provider directly (such as Linode Tokyo[15]) and the GFW is able to analysis network traffic patterns and block certain ports of the server, such as 22 (SSH).

3. VPN

   A virtual private network (VPN) is the extension of a private network that encompasses links across shared or public networks like the Internet. A VPN enables users to send data between two computers across a shared or public internetwork in a manner that emulates the properties of point-to-point private link[16]. Its often used in private network environments such as a company, or a school for employees/students to access the data stored or servers in the intranet. It provides a encrypted and safe way for authorized users to connect to a private network. There are three different protocols for data transferring within a VPN, including Point-to-Point Tunneling Protocol (PPTP), Layer Two Tunneling Protocol (L2TP) and OpenVPN. And L2TP provides stronger authentication and encryption (256-bit vs 128-bit) so that its more safer than PPTP while PPTP is easier to deploy. OpenVPN is an open-source

---

[13] ShadowSocks by Clowwindy. https://github.com/clowwindy/shadowsocks

[14] GoAgent. https://code.google.com/p/goagent/ (Chinese)

[15] Github Blocked in China. http://www.reddit.com/comments/16zs4u

[16] Virtual Private Networking: An Overview. http://technet.microsoft.com/en-us/library/bb742566.aspx

implementation of VPN techniques. It provides good encryption and requires less CPU power than L2TP.

There are many free and charged VPN service providers on the Internet, and most of the free VPNs have been blocked due to their popularity. Charged VPNs are safer but still cannot guarantee that they are accessible all the time. Generally, L2TP VPNs are safer to use than PPTP VPNs.

4. Others
   There are also other assistant methods to bypass the block, such as changing DNS server / DNSCrypt and special browsers on mobile platforms.

8. A Simple Proxy Server[17]

A simple HTTP proxy can be easily developed with Python and some understanding of unix network programming. Basically it accepts HTTP/HTTPs requests from sockets, get the addresses of the requests, connect to remote servers and transfers data between the client and the remote server.

```
# Server Start
def start_server():
    soc = socket.socket(socket.AF_INET)
    soc.bind(('localhost', 8080))
    soc.listen(0)

    while 1:
        # wait for connections forever
        thread.start_new_thread(ConnectionHandler, soc.accept()
+(60,))

if __name__ == '__main__':
    start_server()
```

The `start_server` method is very easy to understand, it sets up a socket for IPv4 address families, and go into a infinite loop accepting connections. `soc.accept()` will return a pair `(conn, address)` where conn is a net socket object usable to send and receive data on the connection[18]. So when `soc.accept()` returns, a new thread will be created and Connection Handler will handle the connection in a separate thread (non-block).

---

[17] Based on https://code.google.com/p/python-proxy/source/browse/trunk/PythonProxy.py

[18] Python Docs. http://docs.python.org/2/library/socket.html

```
class ConnectionHandler:
    def __init__(self, connection, address, timeout):
        self.client = connection
        self.client_buffer = ''
        self.timeout = timeout
        self.method, self.path, self.protocol =
self.get_base_header()
        if self.method=='CONNECT':
            self.method_CONNECT()

        elif self.method in ('OPTIONS', 'GET', 'HEAD', 'POST',
'PUT', 'DELETE', 'TRACE'):
            self.method_others()
        self.client.close()
        self.target.close()
```

The constructor of connection handler initializes some properties, and depend on the method of the connection (retrieved by `self.get_base_header()`), establish the remote connection in different ways by calling `self.method_connect()` and `self.method_others()`.

```
def get_base_header(self):
    while 1:
        self.client_buffer += self.client.recv(BUFLEN)
        end = self.client_buffer.find('\n')
        if end!=-1:
            break
    # method, path and protocol
    data = (self.client_buffer[:end+1]).split()
    # store http header for latter connection
    self.client_buffer = self.client_buffer[end+1:]

    return data
```

The get_base_header method first receives the HTTP method[19] (first line) and HTTP header[20] from the client. The HTTP method is the first line of the content and HTTP header is the rest, and the variable "`end`" stores the position of the first line separator in the content.

[19] HTTP Methods. http://www.w3.org/Protocols/rfc2616/rfc2616-sec9.html

[20] HTTP Header Fields. http://www.w3.org/Protocols/rfc2616/rfc2616-sec14.html

```
GET http://www.google.com/ HTTP/1.1
Host: www.google.com
Proxy-Connection: keep-alive
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/
*;q=0.8
User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10_8_3)
AppleWebKit/537.31 (KHTML, like Gecko) Chrome/26.0.1410.65
Safari/537.31
X-Chrome-Variations: CM21yQEIlLbJAQibtskBCKK2yQEIqLbJAQiptskBCK
+2yQEI94PKAQiGhMoB
Accept-Encoding: gzip,deflate,sdch
Accept-Language: en-US,en;q=0.8
Accept-Charset: GBK,utf-8;q=0.7,*;q=0.3
Cookie: ...
```

Typical base header

So what get_base_header does is split the first line of the content sent by the client to get the method, path of protocol of the connection, and store the HTTP header fields in `self.client_buffer`.

```
def method_connect(self):
    try:
        self._connect_target(self.path)
        self.client.send('HTTP/1.1 200 Connection established
\n')
        self.client_buffer = ''
        self._read_write()
    except:
        print 'Error connecting to remote server'

def method_others(self):
    self.path = self.path[7:]
    i = self.path.find('/')
    host = self.path[:i]
    path = self.path[i:]
    try:
        self._connect_target(host)
        self.target.send('%s %s %s\n'%(self.method, path,
self.protocol)+ self.client_buffer)
        self.client_buffer = ''
        self._read_write()
    except:
        print 'Error connecting to remote server'
```

The method_connect method is used to make connections. It first connects to the remote address retrieved by self.get_base_header and stored in self.path. And if succeed, send a connection

succeed status code[21] back to the client.

On the other hand, other HTTP method requests such as GET and POST will be handled by the `method_others` method. It first retrieves the hostname and resource path from `self.path`, connect to the host and then reconstruct the base header and send the data to the remote server.

```
def _connect_target(self, host):
    i = host.find(':')
    if i!=-1:
        port = int(host[i+1:])
        host = host[:i]
    else:
        port = 80
    (soc_family, _, _, _, address) = socket.getaddrinfo(host,
port)[0]
    self.target = socket.socket(soc_family)
    self.target.connect(address)
```

The connection to the remote server is established through a private method `_connect_target`. The `socket.getaddrinfo` call translates host and port into a sequence of 5-tuples `(family, sockettype, proto, canonname, socketaddr)` that contain all the necessary arguments for creating a socket connected to that service.[22] Then the method creates another socket based on the information returned by `getdddrinfo` and stores the connection in `self.target` for latter use.

```
def _read_write(self):
    time_out_max = self.timeout/3
    socs = [self.client, self.target]
    count = 0
    while 1:
        count += 1
        (recv, _, error) = select.select(socs, [], socs, 3)
        if recv:
            for in_ in recv:
                data = in_.recv(BUFLEN)
                if in_ is self.client:
                    out = self.target
                else:
                    out = self.client
                if data:
                    out.send(data)
                    count = 0
        if count == time_out_max:
            break
```

[21] HTTP Response Codes. http://www.w3.org/Protocols/rfc2616/rfc2616-sec6.html

[22] Python Docs. http://docs.python.org/2/library/socket.html

The final method in the ConnectionHandler class is _read_write. It sends/receives data from/to the server/client and its the heart of the proxy tunnel.

The implementation of the method is relatively simple. Its a infinite loop that transfers data between the client and the target server. And if it failed to receive any data within the timeout, the loop will break and the connections will be closed by the constructor method. `select` is a standard unix system call that wait objects for reading, writing or exception condition. It returns when a triple of lists of objects that are ready[23].

The server can be easily tested locally by running the python script and then setting the HTTP proxy of the browser/system to localhost:8080.

9. Conclusion

In conclusion, the GFW is a invisible wall that blocks people in mainland China from the Internet. Its a tool for abused by Chinese governors and a tool for control. Its a sign of unconfident in the government - that they can only block the content they don't want to see. And one of the worst outcome is that its forcing people censor themselves - a very dangerous sign. People should no matter what go against tools such as the GFW.

[23] Python Docs. http://docs.python.org/2/library/select.html#select.select

References

Anonymous. Finding the contributors to Great Firewall by their papers. https://docs.google.com/a/murraystate.edu/document/d/1mReoa-dEVhLTZ9agyBTLueUX8Fd0er7BaV24iMJYknI/edit?hl=en&forcehl=1

Anonymous. The collateral damage of internet censorship by DNS injection. http://www.sigcomm.org/ccr/papers/2012/July/2317307.2317311.

China's blocking Facebook and Twitter. http://techcrunch.com/2009/07/07/china-blocks-access-to-twitter-facebook-after-riots/. Retrieved April 19, 2013.

Xiong Gang, Meng Jiao, Cao Zi-gang, Wang Yong, Guo Li, Fang Binxing, Research Progress and Prospects of Network Traffic Classification. Journal of Integration Technology, Vol 1, May, 2011

Overview of the Public Security Sector, Cisco Systems, Inc. http://www.wired.com/images_blogs/threatlevel/files/cisco_presentation.pdf. Retrieved April 21, 2013.

Empirical Analysis of Internet Filter in China. http://cyber.law.harvard.edu/filtering/china/appendix-tech.html. Retrieved April 21, 2013.

Fang Binxing: Research Progress and Prospects of Network Traffic Classification: http://jcjs.siat.ac.cn/ch/reader/view_abstract.aspx?file_no=201205006&flag=1

Github SLL replaced by self-signed certificate in China: https://news.ycombinator.com/item?id=5124784. Retrieved April 21, 2013.

Richard Clayton, Steven J. Murdoch, and Robert N. M. Watson, Ignoring the Great Firewall of China. http://www.cl.cam.ac.uk/~rnc1/ignoring.pdf

Tor partially blocked by China. https://blog.torproject.org/blog/tor-partially-blocked-china. Retrieved April 21, 2013.

Picturing Tor censorship in China. https://blog.torproject.org/blog/picturing-tor-censorship-china. Retrieved April 21, 2013.

Xueyang Xu, Z. Morley Map, and J. Alex Halderman, Internet Censorship in China: Where Does the Filtering Occur? Department of Computer Science and Engineering, University of Michigan. http://pam2011.gatech.edu/papers/pam2011--Xu.pdf

RFC 793. Transmission Control Protocol: Protocol Specification. http://www.ietf.org/rfc/rfc793.txt. Retrieved April 21, 2013.

RFC 2616. Hypertext Transfer Protocol - HTTP/1.1. http://www.ietf.org/rfc/rfc2616.txt. Retrieved April 21, 2013.