Hacking the Great Firewall

Mitchell Edwards
Sponsored: Dr. Phillip Rhodes

Senior Project, Spring 2018

Abstract

Censorship is defined by the Oxford Dictionary as the suppression or prohibition of any parts of books, films, news, etc. that are considered obscene, politically unacceptable, or a threat to security.[1] The definition differs in the modern era, because as times have changed, censorship has become both harder than ever and significantly easier, both blatantly obvious and incredibly easy to hide. The era of the internet brought about a much more public forum to discuss what couldn't be discussed before, and by-and-large the Western leaders of the free world have encouraged free-thinking, online debate, and progressive discussions online, with very few caveats.

There is, however, an Eastern power, like the Mordor of the internet, that threatens this free-speaking haven of the internet. China has, with bold pride, established the most extensive instrument of suppression in history. The Great Firewall of China has stood for as long as the Chinese internet, growing stronger every year as the technology boom both feeds and destroys innovation. This censorship mechanism exists as far more than an internet blacklist of terms, sites, and searches. It often takes the form of swarms of patriotic web-warriors, flooding a post thought to be politically incorrect and bashing the writer. Sometimes, it is a notification in your inbox that your post has violated guidelines. Others, that notification takes the form of a knock on your door.

The 'Hacking the Great Firewall' project, formerly known as the 'Hopping the Great Firewall' project, began as an attempt to peer behind the wall and quantify the mechanism we all knew existed yet knew little about. I sought to understand just what was taboo and what was accepted, what volume of posts were being formally taken offline and what were allowed through. The project morphed over time into a virtual cat-and-mouse game, a back and forth struggle between one of the world's most advanced dictators of online media, and a soon-to-be college graduate with questionable web-scraping abilities and a lot of energy drinks.

[1] https://en.oxforddictionaries.com/definition/us/censorship

While this seems like an action movie mixed with a cypherpunk dystopian novel, instead it was a series of lessons learned, set-backs, and insights. I changed my way of looking at the project mid-way through, changing it from a quantification of Chinese censorship to an insightful presentation of taboo discussion. Technically, the project morphed from a basic web-scraper to a hacked together anti-censorship mechanism in and of itself, with the aim of not just browsing material, but archiving it for future use and digestion.

What I present at the end of the road is both less than and more than I hoped. Technically, the Hacking the Great Firewall project is one of the more complex systems I have worked with. Philosophically, it is an internal look into the psyche of a half-man, half-technology machine that ingests millions of posts per day and siphons through them, searching for a dynamic definition of political incorrectness. Culturally, it is a better understanding of the Chinese mindset towards media, both a look at the antagonistic dissidents and the harmonious tyrants. I offer a study of the data I've gathered, comparatively tiny in size, but more valuable than gold in insight.

Finally, I would like to reiterate one of the most important points in the completion of this project. The HtGFW project is published under an Open Source License, meaning the source code from the beginning of development forward is entirely open source. I plan for it to stay that way. No user data was gathered for the project aside from user handles, so the data itself is not sensitive and will be kept Open Source as well. I intended from the beginning of development to make this project accessible to anybody who wanted to continue development, analyze the open data source, or otherwise consume the source code or data, and I believe keeping the data and the source code open and public is important for that aim.

The purpose of this project was to gain insight into the effect the Great Firewall has on the Chinese internet. I discovered more than I intended and have a much greater understanding of the censorship apparatus than when I began development.

Hacking the Great Firewall: A Game of Cat-and-Mouse

Introduction

The Hacking the Great Firewall (HtGFW) project was an incredibly educational journey toward a better understanding of the massive censorship apparatus in China. I had a relatively deep understanding of how the Great Firewall *might* work, but no quantified evidence to back up the conjecture. The Great Firewall apparatus proved to be a far more manual, focused project than I thought before, with most of the automated censorship occurring on the "front end," or before the post even appears on the platform.

The project was focused on the Chinese language social media platform Zhihu, a close cousin to Yahoo answers. The platform is divided into various topics across a wide spectrum, including politics (政治) , social issues (社会问题), economics (经济), and democracy (民主). The project mainly sought to discern which topics were under the most scrutiny, which topics were censored over differing time periods, and what kind of cultural and political background influenced censorship of different topics.

Hacking the Great Firewall took many forms over time, beginning as a singular iteration of a basic python scraper, evolving out of necessity into a distributed product on three different Virtual Private Servers (VPS). It began as a list of ten, likely to be censored topics, growing into a list of double that with recommendations from users on social media. This necessity, ironically, came from a series of battles with the censors themselves, battles fought in the form of CAPTCHAs, IP/MAC bans, and strange Deep Packet Inspection (DPI).

Final Product

The final product, a result of a series of challenges and trials both inherent in any develop process and inherent in any challenge to a major censorship regime, resides on three remote Virtual Private Servers running on Ubuntu 16.04. The project was written entirely in Python, and consists of three different pieces: the Zhihu API, the Dumping module, and the Checking module.

The Zhihu API forms the core of the project, and had to be built from scratch for this project. The functionality is relatively basic, allowing the user to dump ten or so posts from each topic. Topics are represented by a unique, numerical designator, which the API can also translate into a topic name, ie 20073350 -> 共产主义（communism).

The Dumping module uses the Zhihu API to dump topics based on a user-supplied text file populated with the numerical topic designators. They save meta-data of the posts, including the topic designator, post designator, poster username, and the link to the post, as well as the post title and post body.

The Checker module loops through the posts at different intervals to check the urls for censorship. If the post is censored, the text file is then archived in a different directory ('./Censored/topic#').

The whole system is automated via simple bash scripts that run the Dumping and Checker module at different time intervals. These time intervals were tweaked to avoid detection by Zhihu system administrators. These time intervals were a key part of balancing data volume with detection by censors.

Analysis

At last count, the HtGFW project gathered 5,417 posts, with 64 posts being detected as being censored. The top 5 censored topics are as follows: Societal Issues (社会问题), Politics (政治), Freedom (自由), Society (社会), and Media (媒体).

Before development began, I hypothesized that political discussion would be the most censored topic, with Taiwan and Hong Kong centric discussion being particularly censored. I also believed that censorship mainly consisted of automated flagging based on keywords of posts already on the site.

By the end of the project, I discovered that societal issues, topics such as rampant student suicide rates, the democratic party of Taiwan (国民党), and differing subjects surrounding cryptocurrency and topics that seem to oppose Chinese cultural norms. Discussion about Taiwan, Japan, and

India were also heavily censored, but not as much as the aforementioned cultural and societal topics.

Another surprising discovery was that most of the censorship occurred on the front end. I noted early on that the volume of censored posts was noticeably low. By the end of the project, only about 1.5% of posts scraped from the platform were removed. Upon manual investigation, posts containing words like Kuomintang (国民党, the Democratic Party of Taiwan) were censored before they were ever posted. This represents a much more severe form of censorship, one that stifles free speech before it ever has a chance to occur.

The case of Yang Baode (杨宝德) is a pertinent case study. Many of the top censored posts were discussions of Baode's suicide that allegedly resulted from sexual harassment by his college TA. The case appears to be a result of corruption in the academia leading college campuses in China, with allegations of rampant coverups of the case being a hot-topic on Chinese social media. This represents a pertinent link between social issues like sexual assault and harassment and political issues surrounding leading Communist Party officials in Chinese academia.

The discovery of heavy censorship of social issues represents an important paradigm shift in the way the West views censorship of Chinese social media. Obviously, politics are a no-go issue online, but the societal focused censorship mechanism shows that the Great Firewall does not stop at politics. Censorship of social issues represents a system centered around societal repression that seeks to stifle not just hot-button political topics but progressivism in general.

Challenges

From the beginning of development, I was fighting an uphill battle against the very system I sought to analyze. The site was difficult to scrape, with an extremely layered design that made digging through lines of HTML tiresome. The site layout changed once or twice after I began scraping, whether by coincidence or purposefully forcing me to change the platform.

Perhaps the largest challenge from the beginning of development was the installation of CAPTCHA forms. CAPTCHA forms, by design, make it much more difficult and significantly slower to scrape data from a site, as it forces either a manual or automated CAPTCHA solving algorithm or a manual slow-down of the scraping platform to escape the CAPTCHA mechanism by appearing closer to legitimate traffic.

It must be stated that the CAPTCHA forms were not in place when I started scraping. The beginning algorithm had no slow-down functionality in place, scraping data as fast as the requests could return. They were a new development and seemed to change their behavior during the course of the development. For example, in the beginning, there was a one-second sleep call between each request. At that point, that was all that was needed, and the scraping module worked for a while. At some point, though, the scraper kept hitting the CAPTCHA wall, forcing me to increase the sleep call so as to avoid the CAPTCHAs. At one point, the CAPTCHA did not even offer a picture or input form, functionally blocking the program with no means of solving the CAPTCHA. I hypothesize that at this point they (system administrators of Zhihu) believed I was using some kind of CAPTCHA solving algorithm, not knowing that my solution was far more rudimentary.

There came a point as well where all requests would return a 404 status code, fooling my platform into believing that all posts had been censored or the platform had been blocked, even though navigating to the web pages using a standard browser, under the same IP, returned the original post. I never quite figured that problem out, and I believe I fixed it when I changed my header and cookie fingerprint.

Fighting the uphill battle taught me more about the censorship mechanism than the platform itself. The use of header and cookie screening to block the platform proved that the Zhihu system administrators were employing some level of widespread Deep Packet Inspection (DPI), and the constant IP/MAC bans and CAPTCHA changes confirmed that they didn't like someone peering over the Wall. I believe that the data gathered by the platform is incredibly valuable, but pales in comparison to the clear and prescient insistence of manually shutting down free speech and oversight.

Looking Forward

Going forward, I hope to keep the platform running on the DigitalOcean servers. I've set up a Patreon page to help crowdfund the hosting costs and plan to tweak the platform to make it more efficient. If I can support it financially, I hope to expand both the gathering capabilities and the analysis functionality to add a larger data set and a more robust analysis tool to garner more intelligence from the data I gather. I don't have a clear vision for how the project will expand, but judging by the initial numbers, I plan to eliminate some topics that don't seem to garner much censorship in favor for others that potentially might. I also would like to add a keyword search functionality, both to check to see if a word is censored on the front-end of the platform as well as to check existing data on censored posts to see if the keyword comes under particular scrutiny.

This is a project I am incredibly passionate about, and have been for a long time. I look forward to the continuing development, refinement, and expansion of the Hacking the Great Firewall platform, and hope that the intelligence gathered from the data will have even a small part in ensuring a freer internet for the citizens of China.