Chinese Online Censorship

Wendy Xue, Shaohan Chen, and Deb Linton

Introduction

It's common knowledge that China censors it's internet — changing the way it's citizens access information and view the world. However, researchers are still trying to understand the anatomy of this censorship so that better international policies may be implemented.

Our research team partnered with Professor Xiao at U.C. Berkeley's School of Information who has been amassing datasets that capture the impact of Chinese censorship. Our aim was to develop an interactive dashboard to help users better understand the types of sites were being censored and the methods of censorship China uses. We also hope this dashboard might be of use to the original researchers on Professor Xiao's team. Their ongoing work involves routinely assessing large data samples that might be easier to review in this visual format.

Questions our dashboard attempts to answer:

+ What websites are being censoring by China?
+ How aggressively are those sites being censored?
+ Are there any geopolitical patterns to what domains are censored?
+ Do certain types of sites get censored more frequently or more aggressively?
+ How might Chinese internet culture differ from global internet culture?

Discussion of related work

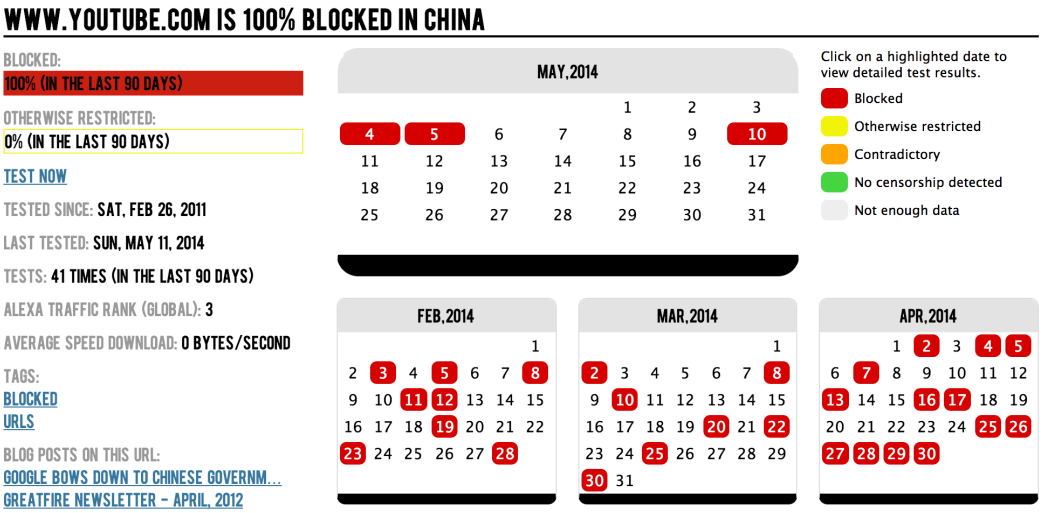
There exists several websites that present Chinese online censorship in visual format. We evaluated these visualizations in terms of their ability to convey meaning and the use of visual components. Reviewing this work helped inspire us to design our own visualizations and avoid their pitfalls.

1. Visualizations on GreatFire.org (https://en.greatfire.org/)

This website contains two types of visualizations of popular domain names that are blocked by Chinese censorship. The first type of visualization is a wordcloud of popular blocked domains as shown below.

The advantage of using wordcloud to illustrate the blocked domains is that viewers could easily spot a few most popular domain names. However, it is not effective to provide the big picture of the censorship landscape. There are several problems with the wordcloud design. First, viewers tend to focus on the biggest domain names and neglect less popular domains. This is problematic because less popular websites could be highly important websites, such as activist blogs. Second, it is hard to compare the text size of the domains, especially when some are placed horizontally and some are vertically. For example, it is difficult to tell whether twitter.com is more popular than youtube.com. Third, a wordcloud does not tell viewer by how much one domain is more popular than another. While frequency can show us some patterns of censorship, the wordcloud alone does not tell viewer what type of contents these blocked domains contain. Overall, the wordcloud does not provide viewers any interactive means of exploring the data.

The second type of visualization on the GreatFire.org website is the calendar view of the results of testing conducted over time to verify whether a website has been censored. The following figure shows the censorship calendar using Youtube.com as an example.



We found the calendar view quite compelling in demonstrating how long censorship has been in effect for a particular website. The choice of color makes it easy for viewers to tell whether the website was blocked on the day of the test. We felt that the calendar view would be a great option to use if we could obtain a time series of data. Unfortunately, Professor Xiao's research team did not have data over a period that was long enough to make a time series visualization.

2. Visualizations on Information Is Beautiful (http://www.informationisbeautiful.net/visualizations/what-does-china-censor-online/)

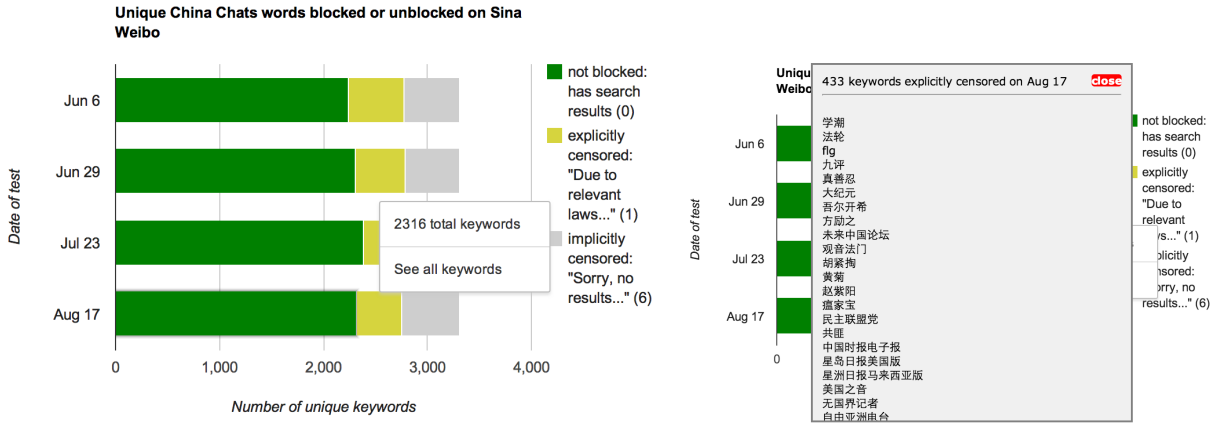
We found an interesting visualization, shown below, on informationisbeautiful.net, presenting censored contents by using geography shape.

The blocked search keywords are colored in red and make up the shape of China, whereas the blocked domain names are shown in black and are listed outside of the China shape. This arrangement suggests that the blocked domains are inaccessible from inside of China. We thought the use of call-outs for explanations were clever attempt to help viewers understand why certain keywords are blocked. Nevertheless, the arrows do not point to the correct terms and therefore have a misleading effect. For example, the word “Xinjiang” on second bottom call-out is linked directly to the domain “tibet.com”. The intended keyword to be linked is actually the third word from the right on the same row. The biggest difficulty we had for this visualization was that it is unclear that why certain keywords have explanation and others don’t. Besides, there isn’t any obvious categorization or ordering of keyword. For viewers who are unfamiliar with how Chinese online censorship works, they might not understand that the words in red are blocked search keywords.

3. Visualization on blockedweibo.tumblr.com (http://blockedonweibo.tumblr.com/post/58933956184/interactive-charts-showing-changes-in-weibo-keyword)

A third website we searched and found visualization about China’s online censorship showed an interactive stacked bar chart of blocked versus unblocked search keywords on Sina Weibo, a

Chinese social media platform similar to Twitter.com. As shown in the figure on the left below, as viewers hover over the stacked bars, a tooltip window is displayed with a count of number of keywords in the corresponding category, and a link to see the list of keywords in that category. The figure on the right shows how the list of keywords look like after clicking on the link.



Among other visualizations we have discussed above, we felt that this bar chart was the weakest in terms of communicating the meaning of data effectively. Although it breaks down the blocked content by the types of search responses returned by Sina Weibo, it lacks information about unique characteristics of each blocked keywords. Viewers are also left wondering what was the purpose of the stacked bar chart. Is it to demonstrate how large or how small the percentage of keywords being blocks is? We felt that the static inforgraphics in previous sections convey more meaning than the bar chart here, even though the bar chart is interactive.

Visualization Design

After reviewing other related works above, we then started designing the visualization and user interactions. Since we would like to show the fact of “how websites are blocked by China government around the world” and “the impact of top websites that are blocked in China”, we decided to use two main infographics to show our data - a world map to show blocked sites’ geolocations derived from IP addresses and blocking methods, and a zoomable tree map to show the traffic of top 150 blocked websites and the numbers of these blocked websites under each category.

World Map

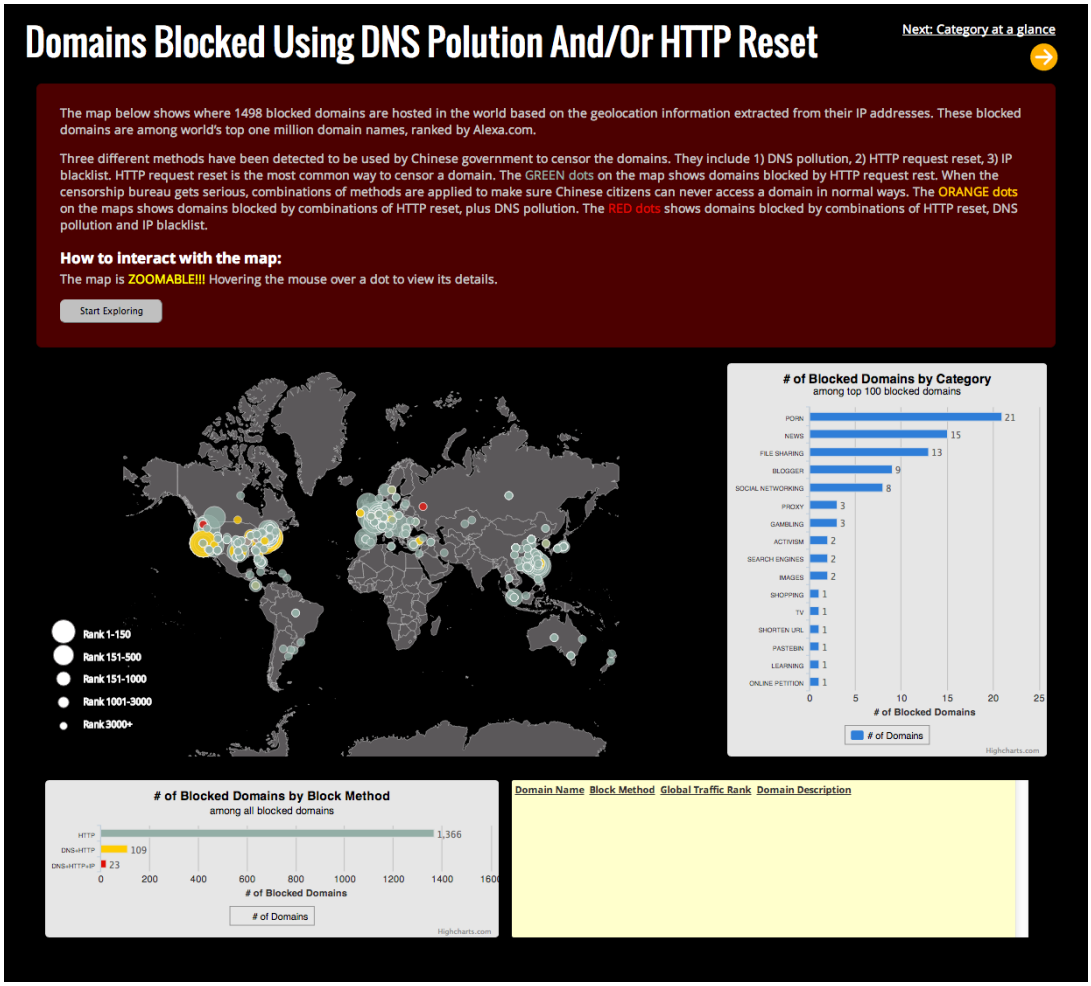
In the world map page, we used dots of various colors and sizes to represent essential information about these blocked websites.

- **Location:** The location of the dot indicates the geolocation of its IP address
- **Size:** The size of the dot indicates it's ranking (by traffic) as calculated by Alexa.com. The larger the radius is, the higher the website is ranked
- **Color:** The color of the dot indicates the method(s) that is being used to block that website. There are three different colors in this map, which represent different combinations of blocking methods.

On the dashboard, there are also two bar charts showing the category and methods being used to block the sites. We first came out this idea because we realized there were too many dots on the map and was kind of hard to catch the rough idea in numbers, and bar charts are excellent for presenting quantified data. After several design narratives, we found that bar charts can actually work as filter that users can just click a specific bar to view related data on the map. As a result, we decided to implement bar charts with filter feature into the dashboard, and use colors to associate blocked methods both on map and in the bar charts.

Besides, since some users may also want to know which website a dot represent, we implemented the tooltip display feature on the map as well. When user hovers on a specific dot, the essential data such as name of the website and method(s) being used to block this website will be shown. When users click on a bar inside the bar chart, the list of blocked websites under this specific category will also be shown on the screen.

(image: draft of world map and treemap dashboard)



(image: final design of world map dashboard)

Treemap

In the treemap view, we implemented a filter for users to view the data in different ways - by number of sites and by website traffic. Under “number of sites” view, each block is divided equally under each category, and user can easily see the categories that have the most number of blocked sites. Under “website traffic” view, however, each block is divided by its traffic. The more traffic a website has, the larger area the block will be.

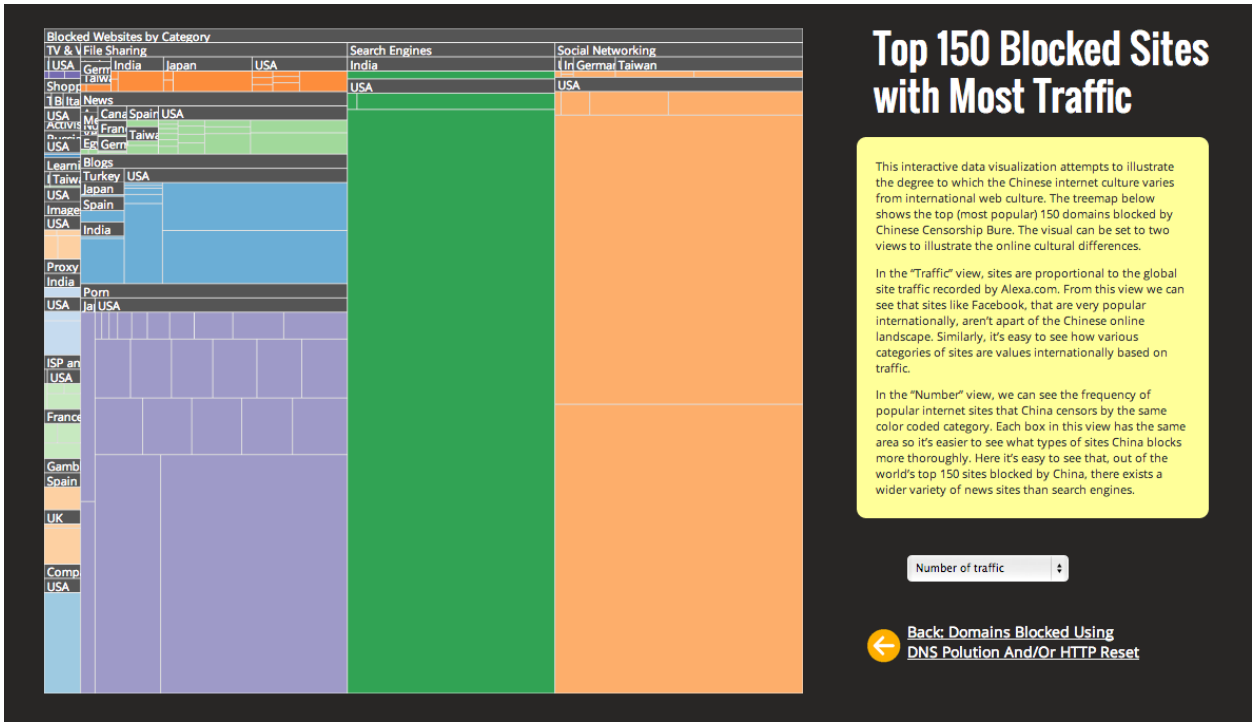
The reason we setup two different filter is to let user interact with the data and compare the differences by using animated visualization. For example, when user selects “number of sites” view, they can see that there are only four websites under “Search Engine” category and the category shares just a tiny portion in this treemap. However, when user switches to “traffic” view,

he or she will find that the “Search Engine” category shares 2nd large portion in the treemap. This is because some of the blocked search engine websites are world’s top websites by traffic, and it also indicates that blocking these websites have higher impact than blocking others.

Below are the essential elements of the treemap:

- **Size:** In traffic view, the size of a block indicates the traffic volume of the website. The more traffic of a website has, the larger the block area will be. In site number view, each block under the same category is divided equally.
- **Color:** The color in both number of sites and sites traffic view indicates category. Websites under same category will be the same color.
- **Header:** There are three different levels of layer in this treemap. The top level shows the title of this treemap, the second level shows the category, and the third level shows the country these blocked websites locate.
- **Text inside the cell:** Text inside each color cell indicates the URL of the represented site. When user click till the bottom layer, the URL will show up that user can refer to.

There was a debate about the color during our initial design. We originally used color to indicate country, which in other words means that websites under same country will have same color. The original idea was to associate this treemap with world map that users can easily understand which country have the most blocked websites in the top 150 sites. However, after our first implementation, we realized that using colors to indicates country actually didn’t make too much sense, since there are way too many countries on the treemap that the whole map looked too colorful and lost focus. On the other hand, since the main reason for creating this treemap was to show the impact of blocked websites under different category, we then decided to assign the color based on category rather than by country.



(image: final design of treemap dashboard)

Dataset and Data Preparation

We obtained two sets of data from Professor Xiao's research group. One of them was a large dataset containing 18000 blocked domain names with their sub domain names, IP addresses, blocked methods, popularity rank from Alexa.com, and latitude/longtitude values for each IP. The second dataset was a list of top 150 (ranked by popularity) blocked domains with categories assigned manually by the research group. The first dataset was broken into four CSV files. Below is a sample of the data.

Domain	Rank	SLD	SLD Rank	DNS	HTTP	IP	IP Details				
							IP	Blocked	Country	City	Lat/Long
facebook.cor	2	facebook.com	2	Y	Y	Y	173.252.110.27	Y	US	Menlo Park	37.459/-122.1781
youtube.com	3	youtube.com	3	Y	Y	N	74.125.235.142	N	US	Mountain View	37.4192/-122.0574
							74.125.235.128	N	US	Mountain View	37.4192/-122.0574
							74.125.235.129	N	US	Mountain View	37.4192/-122.0574
							74.125.235.130	N	US	Mountain View	37.4192/-122.0574
							74.125.235.131	N	US	Mountain View	37.4192/-122.0574
							74.125.235.132	N	US	Mountain View	37.4192/-122.0574
							74.125.235.133	N	US	Mountain View	37.4192/-122.0574
							74.125.235.134	N	US	Mountain View	37.4192/-122.0574
							74.125.235.135	N	US	Mountain View	37.4192/-122.0574
							74.125.235.136	N	US	Mountain View	37.4192/-122.0574
							74.125.235.137	N	US	Mountain View	37.4192/-122.0574
twitter.com	8	twitter.com	8	Y	Y	Y	199.59.148.10	Y	US	San Francisco	37.7697/-122.3933
							199.59.148.82	Y	US	San Francisco	37.7697/-122.3933
							199.59.149.198	Y	US	San Francisco	37.7697/-122.3933
blogspot.con	15	blogspot.com	15	N	Y	N	74.125.128.191	N	US	Mountain View	37.4192/-122.0574
wordpress.cc	18	wordpress.com	18	N	Y	N	66.155.11.243	N	US	San Francisco	37.7484/-122.4156
							192.0.82.252	N	US	San Francisco	37.7484/-122.4156
							76.74.254.126	N	US	San Antonio	29.4889/-98.3987
papapacc.tumblr.com		tumblr.com	35	N	Y	N	66.6.40.38	N	US	New York	40.7391/-73.9826
							66.6.40.58	N	US	New York	40.7391/-73.9826
www.xvideos.com		xvideos.com	43	N	N	Y	69.55.52.253	Y	US	New York	40.7267/-73.9981
xvideos.com	43	xvideos.com	43	N	Y	Y	69.55.52.253	Y	US	New York	40.7267/-73.9981
xhamster.cor	56	xhamster.com	56	N	Y	N	88.208.24.57	N	NL	Amsterdam	52.35/4.9167
							88.208.24.58	N	NL	Amsterdam	52.35/4.9167
							88.208.24.56	N	NL	Amsterdam	52.35/4.9167
							88.208.24.59	N	NL	Amsterdam	52.35/4.9167
it.xhamster.com		xhamster.com	56	N	Y	N	88.208.24.59	N	NL	Amsterdam	52.35/4.9167
							88.208.24.58	N	NL	Amsterdam	52.35/4.9167
							88.208.24.56	N	NL	Amsterdam	52.35/4.9167
							88.208.24.57	N	NL	Amsterdam	52.35/4.9167

(image: sample dataset)

We decided to plot the domain names by the location of their hosting IP addresses onto a world map. To prepare the data, we created a Python script to clean the large domain name dataset. We first merged the four CSV files into one file, and then excluded any domain that was blocked only through IP blacklist. This was because the test used by the research group to verify if a domain was blocked by IP blacklist produced unrealiable results. There were domains were found to have several hosting IP addresses with some of them on the blacklist while some not. Therefore, we decided to include domains that were blocked by at least HTTP reset or DNS pollution, as the research group confirmed that results for these blocking methods were for sure correct.

We also merged redundant domain data to further reduce noises in the records. In the raw dataset, domain names with prefix “www.” were listed in addition to their counterparts that without the prefix. Their sub level domain names were the same. Occasionally, the domain with the prefix was listed as blocked by DNS pollution, and the one without prefix was listed as blocked by HTTP reset. We merged these domain data to create one record using the sub level domain name. The blocking method becomes a union of the methods from previous records. The cleaned dataset was converted to a CSV file.

We felt that it was important to show viewers the description of a domain when they explore the data using the world map. As the original dataset did not contain domain description, we wrote a Python web crawler to scrape information from Alexa.com, which contains a website description section on the domain information page. If the description was found, we appended it to our dataset. During a period of time, our web crawler had accessed Alexa.com so frequently that Alexa blocked our IP. We had to wait for a few days before we were able to continue crawling the data.

In order to draw the bar charts, we created several Python scripts to do the math and convert the data into JSON format. The dataset with manually created categories required cleaning before conversion as well. There were several spelling mistakes in the category names, as well as missing categories. We had to manually visit those domains with missing categories and fill in the data. Furthermore, we needed to parse the domain names from the category dataset, because the domain names, the popularity rank, and their listing order were contained in the same string. Once the dataset was cleaned, we converted it into a hierarchical JSON file to be used by the Treemap.

Below is a sample of the category dataset.

Number	URL	Categories	Rank (Alexa.com)	visits over the last 30 days (http://www.trafficestimate.com/)	Country (that represents the biggest percentage of visitor)
	1. 1,google.com,1,DNS|HTTP	Search Engines	1	4.84E+09	USA
	2. 2,facebook.com,2,DNS|HTTP|IP	Social Networking	2	2.67E+09	USA
	3. 9,twitter.com,2,DNS|HTTP|IP	Social Networking	11	6.70E+08	USA
	4. 16,blogspot.com,18083,HTTP|IP	Blogger	16	4.47E+08	USA
	5. 18,wordpress.com,4055,HTTP|IP	Blogger	19	4.04E+08	USA
	6. 39,microsoft.com,1,IP	Companies	39	2.08E+08	USA
	7. 42,xvideos.com,2,HTTP	Porn	42		USA
	8. 60,blogger.com,2,DNS|HTTP	Blogger	58	1.40E+08	USA
	9. 61,fc2.com,1,HTTP	File Sharing	60	1.38E+08	Japan
	10. 65,t.co,2,IP	Social Networking	75	1.19E+08	USA
	11. 77,googleusercontent.com,1,DNS|HTTP	Search Engines	77	1.20E+08	USA
	12. 87,blogspot.in,3462,HTTP	Blogger	89	1.02E+08	India
	13. 102,xnxx.com,1,HTTP	Porn	103		USA
	14. 112,youporn.com,1,HTTP	Porn	109	7.98E+07	USA
	15. 116,dropbox.com,2,HTTP	File sharing	118	8.34E+07	USA
	16. 117,nytimes.com,2,DNS|HTTP|IP	News	117	8.03E+07	USA
	17. 119,slideshare.net,2,HTTP	File sharing	123	7.92E+07	India
	18. 120,pixnet.net,1,HTTP		121	8.09E+07	Taiwan
	19. 142,hootsuite.com,2,DNS|HTTP	Social Networking	141	6.94E+07	USA
	20. 153,soundcloud.com,2,HTTP|IP	File sharing	156	6.29E+07	USA
	21. 164,theguardian.com,2,HTTP	News	162	6.12E+07	USA
	22. 175,bet365.com,2,HTTP	Gambling	188	5.56E+07	Spain
	23. 230,blogspot.com.es,819,HTTP	Blogger	234	4.46E+07	Spain
	24. 329,bloomberg.com,1,DNS|HTTP	News	345	3.49E+07	USA
	25. 331,free.fr,2,IP	ISP	307	3.29E+07	France
	26. 336,mobile01.com,2,HTTP		313	3.50E+07	Taiwan
	27. 340,bluehost.com,1,IP	Web hosting	348	3.36E+07	USA
	28. 354,xing.com,1,HTTP	Social Networking	365	3.11E+07	Germany

(image: sample dataset with manual categorization)

One challenge we had was inconsistency between the two datasets. The category dataset was generated at a different time from the large dataset, thus the popularity rank could be different for the same domain. The category dataset also contained domains that were blocked only by IP blacklist, which were filtered out in the other dataset. Because we planned to make the bar chart functioning as filters of the world map, we decided to align the bar chart data with the world map data. As a result, we only selected the domains that existed in both datasets to produce the category bar chart. The selection was done by creating a Python script to compare the two datasets and generate a new list of domains with categories. This method produced less than 100 domains with categories that were shown on the map.

Implementation Tools

For the implementation, we used the following tools to accomplish our goals:

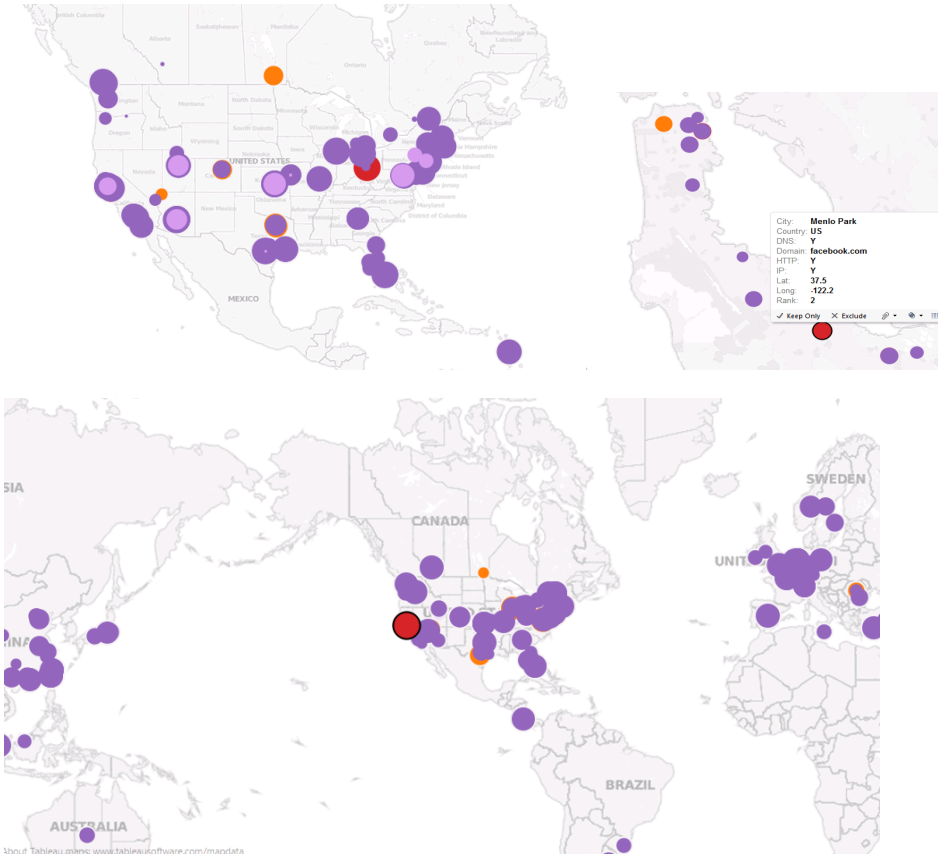
1) **d3.js:** We used d3.js to implement world map as well as tree map. In “Domains Blocked Using DNS Pollution And/Or HTTP Reset“, we downloaded the world shape file from Mike Bostock’s github. The shape file is in topojson format and is generated from natural earth’s medium scale 1:50m world country boundary files. (http://www.naturalearthdata.com/downloads). In “”Top 150 blocked sites with most traffic, we used and modified the template from Mike Bostock’s Zoomable Treemap (http://bost.ocks.org/mike/treemap/) and generated json file based on the recommended structure.

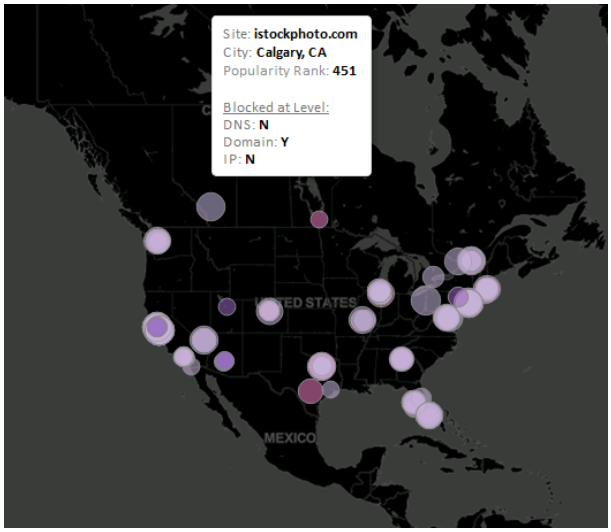
2) **Highcharts.js**: We used Highcharts.js to draw both category and block method bar charts in the world censor map page. These bar charts also act as filters for users to select specific category or blocking method they would like to see on the map.

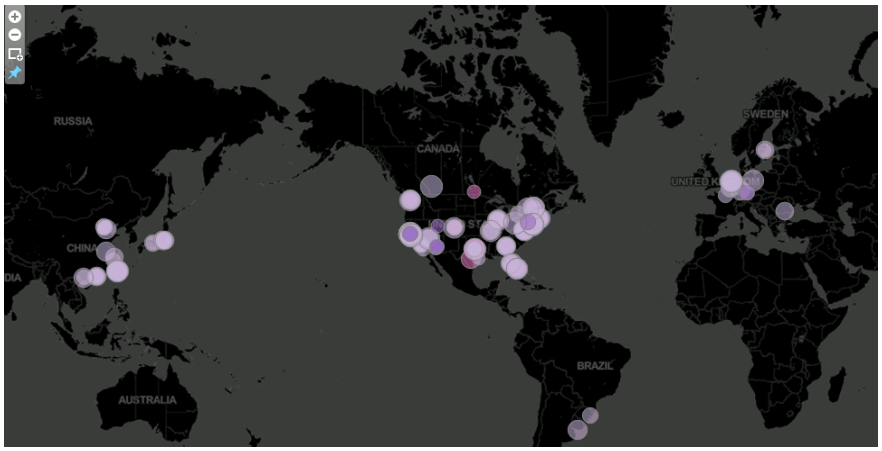
3) HTML/CSS: We used HTML and CSS to assign desired colors and layouts.

Steps were required to accomplish goals

A key part of our design process was using Exploratory Data Analysis (EDA) tools to gain a better understanding of what this data represented and the affordances of various information visualization tools. For us, Tableau was a natural fit given its quick and easy ability to map data and sample color pallets. We also tried CartoDB but our dataset was too large for the basic subscription service. Going through the EDA process helped us identify immediate issues related to clusters of data and the limits to how much information we could code with each data point.







Unexpected Findings

The map of blocked domain names revealed interesting results that were unexpected. We found a cluster of domain names hosted inside of China being blocked by the Chinese government using HTTP request reset method. Our hypothesis had always been that Chinese government censors websites to prevent its citizens to access “unwanted” information that are outside of China. Upon finding domains hosted inside of China being blocked, we took a closer look at these domains and suspected that majority of them are Chinese adult entertainment sites, judging by their names. This finding suggested that the Chinese censor bureau does a sloppy job when the censored contents is not politically sensitive. Although the government does not want Chinese citizens to access adult entertainment websites, when it comes to block it, the

government does not bother to check where the domains are actually hosted. It applies the blocking mechanism blindingly, assuming it will work as normal. If a domain is hosted inside of China, resetting HTTP requests at the border between China's intranet and the outside world will stop traffic going into China. However, it does not have any impact on traffic within China trying to reach the site. Therefore, these "censored" domains are still free to access for Chinese citizens. The figure below shows the blocked domains inside of China.

(image: map showing domains hosted in China)

On the other hand, the data revealed that if a website is deemed as having political sensitive information, or having too much freedom of expression, the government usually go all the way to ensure it's censored aggressively, which means a combination of two or three methods are applied to prevent access to the domain. Examples of these domains include Facebook.com, Twitter.com, Epochtimes.com, and Appledaily.com.tw. The later two are tabloid-style newspapers, often covering news events being censored on China's mainstream media.

A second thing we did not expect to see was the high percentage of adult entertainment sites among the top 100 blocked domains, ranked by web traffic. We have expected the government to censor more news websites and social networking sites. Although there were more aggressively blocked domains in these two categories, there were not as many of them as adult entertainment sites in the top popular domains.

Response from viewers

We evaluated the design with viewers who came to our station during the InfoViz final project showcase. By allowing viewers to explore the visualization first by themselves without us explaining it, we were able to observe their reactions and notice any sign of confusion. When we answered clarification questions, we took notes on what problems viewers had with understanding the visualization. The following is a list of problems we identified by observing and providing clarifications.

1. Several viewers asked about how to read the world map. Users asked about what “Rank” meant, how was the rank determined, and how was the locations of the domains determined. It suggested that we should make explicit explanation. Although we did provide these explanations, the information was burried in the introduction paragraph on top of the world map. Not many viewers had the patience to read through the paragraph.
2. Several viewers didn’t realize the bar charts were clickable and were functioning as filters to the world map data. We should provide help text to the bar chart to make it more obvious. The small size of the bars also made it more difficult for viewers to click on it.
3. Some viewers didn’t realize the table on the bottom corner was populated when data on the world map were filtered. We attributed this problem to the design layout making the table non-obvious.
4. The treemap and the world map toggle was not obvious to viewers and viewers must be prompted to switch the view to treemap or world map. A better layout design could help making the toggle more conspicuous. On the other hand, positioning the treemap at the bottom of the screen might be an alternative.

During the showcase, Professor Xiao and his research team came to our station and we asked them for feedback on the visualization.

1. Xiao’s research team’s feedback were all very positive. They were pleased with the aesthetic appeal of the design. The visualization helped the team to affirm their hypothesis of censorship patterns. It also provided them a visual tool to examine the patterns in further details. The unexpected result about the censored sites in China was a pleasant surprise.
2. The team in general was most impressed with the interactivity between the bar charts and the world map, to which we applied the linking and brushing technique.

3. The team commented on how they liked the consistent color coding of the world map and the bar chart, which made easy for them to understand what the category of data was being examined.

Project Team Contribution

Student Name	Contribution	Percentage
Shaohan Chen	Created treemap using D3.js Created landing page using HTML/CSS Dashboard design and implementation for both world map and tree map pages Authored sections of final report	35%
Deb Linton	EDA, data cleaning, json data for treemap, Authored small sections of the report	25%
Wendy Xue	Data cleaning and conversion. Created world map using D3.js. Created method and category bar charts using HighCharts.js. Implemented interactivity between world map and bar charts. Authored sections of final report.	40%

Appendix

Link to project github repository
https://github.com/shaohan/InfoViz-final-project

Link to project website
http://www.shao-han.com/china-censorship/