Visualizing the Invisible Wall

The impact of the Great Firewall of China on public opinion

Xingyue Li

中国防火墙
微博
推特
公众舆论
Thesis Presented by
Xingyue Li
to The College of Arts, Media & Design
in Partial Fulfillment of the Requirements
for the Degree of Master of Fine Arts
in Information Design & Visualization

Thesis Advisor
Miso Kim

Thesis Reader
Matthew Carroll

Northeastern University Boston,
Massachusetts April 2017
This thesis is a research project visualizing the differences of public opinions in various Internet ecologies using Chinese Internet censorship as an example. The aim of this thesis is to create a visualization showing the shape of China’s so-called “invisible wall,” and its influence on Internet censorship. The target audiences of this visualization are people who are interested in Internet censorship and Internet users both inside and outside China’s firewall. Audiences will have a better understanding of the influence of Internet censorship and generate their opinions by exploring the network visualization and case studies. The visualization is a web-based application so users can share it on social media and get access to the visualization through the web. It introduces the history of Internet censorship and a presentation of case study analyses. The first section is a timeline graph introducing the history of China’s Internet censorship. The second part is an interactive visualization containing topic selection, a tweets scale comparison from each social media platform, account types (whether the tweet is from news outlets or individual users’ posts), and bullet graphs displaying public opinions.
Through the process of thesis research, I have received a lot of help from fellow students and professors in our program. I would like to gratefully acknowledge my thesis advisor, Professor Miso Kim for her consistent support and encouragement.

I would also like to thank all the thesis committee members for the valuable feedbacks and suggestions. I also own gratitude to my thesis reader, professor Matthew Carroll, for his enlightenment and advice. Thanks for Professor Thomas Starr for his tips on the layout of the book. Especially, I would thank Navarjun for the technical assistance on data processing.

Finally, I am grateful to my parents and friends who in one way or another shared their support, either morally, financially and physically, thank you.

Acknowledgement
## Contents

1. Introduction  
   - Objectives 6  
   - Target Audience 6  
   - Methodology 7  

2. Background  
   - Censorship History 12  
   - Internet Censorship 13  
   - History of Chinese Internet Censorship 15  
   - Censorship on Social Media 17  

3. State of the Art  
   - Overview 20  
   - Related Projects 24  

4. Conceptual Framing  
   - Agenda Setting Theory 37  
   - Cultivation 38  
   - Uses and Gratification 39  
   - Influence of Media to Public Opinion 40  
   - Public Opinion on Social Media 42  

5. Experiments Process  
   - First-round Experiment 46  
   - Second-round Experiment 58  

6. Future Work  
   - References 79
The meaning of censor, defined by Merriam-Webster Dictionary, means “a person who supervises conduct and morals such as an official who examines materials (as publications or films) for objectionable matter.” The origin of the word “censorship” can be traced back to Rome in 433 BC. Officials considered censoring information an honorable task and that supervising public morality is vital to “shape the character of the people.”

Throughout history, the objects being censored have included books and scholars during the Qin dynasty (213 BC) in ancient China; religious control of free speech and publication in medieval Europe; press restrictions in the 18th century; mail and packages for military purposes during the American Civil War and both world wars; and Internet content after with the development of information technologies. The censorship in this paper refers to Internet censorship executed by both the government and Internet services companies.

Chinese Internet censorship, also known as The Great Firewall of China (GFW), is a surveillance project for online censorship operated by the government of China. In the beginning, websites were blocked because they mostly contained seditious information that could harm the national security or societal stability in China. Gradually, the scope of those websites being blocked by censorship extended to journalism websites that had divergent political opinions from the government. Then, some popular social platforms and Internet service companies were banned in China. Those websites’ servers are located outside of the mainland and include Google, Facebook, Twitter, and YouTube. In addition, many non-political and academic websites are also included on the government’s “banned” list.


The Great Firewall is one of the world’s most closed network censorship systems developed by a government. Although the government is concerned that controversial information on the web could mislead Chinese Internet users, blocking channels of information will not eradicate information that dissents from China’s official viewpoints. In fact, the sealed-off nature of China’s Internet access could deepen the prejudice and bias between the world both inside and outside of the wall.

An isolated land has been created by the firewall, which makes the networking aspect of the Internet in China vastly limited. It impairs the interconnection of people inside and outside the country. The situation has caused vast disparities in public opinions among Chinese citizens and the rest of the world. With its restrictive Internet technology, the Great Firewall has obstructed the localization of some of the world’s top global Internet service providers like Google, Facebook and Microsoft. The one positive aspect of shutting out these outsized companies is that local companies operate in a market more amenable to their developments and innovations. Chinese Internet users can enjoy excellent Internet services from local companies like Baidu’s search and mapping tools, Youku’s video hosting and streaming, Wechat, QQ, and Weibo’s platform for chatting. To accommodate the government’s information control, these companies have developed their own in-house surveillance systems.

The occurrence of social media makes it possible for different groups of people to directly communicate with each other, which can intensify conflicts and controversies. For example, Chinese netizens “occupied” President Obama’s Google Plus page3 and then “blew up” American pop singer Lady Gaga’s Instagram account after she met with the Dalai Lama4, the exiled Tibetan religious leader. When Internet users living both inside and outside the Great Firewall encounter such conflicts during their online surfing experience, they might not be aware of the existence of the information manipulation mechanisms of China’s Internet service providers and the government’s broader Internet censorship. To show this disparity in information access, it is necessary to show the “shape” of the wall, which reveals the differences in information people can access on both sides of the wall.

The thesis will also present research and visualizations that illustrate the differences of public opinion in various Internet ecologies, using Chinese Internet censorship as an example. The focus of the project is designing methodologies and experiments to develop a data visualization that shows the impact of the Information manipulation. Although the field of information visualization has been developed during the past few years, open data sources and web APIs5 make it possible for designers and journalists to discover valuable information and convey it through visualizations and infographics to the public. There is not yet a project that has grabbed the public’s attention to show the impact of China’s absence of information.

The primary challenge of this project’s data component is accessing and using available information online to reveal the diversity of public opinions on both side of the wall using quantitative science research methodology. With the digitization of journalism, mainstream press websites provide plentiful sources of news formatted as text that can used for analysis. With the development of mobile Internet technology, social media has replaced traditional mass media and portal websites as the preferred venues used by people to acquire information and express their opinions. But these social media platforms have differing reach because of the firewall. As one of the world’s top social platforms with 319 million monthly and active users, Twitter provides services for users outside of mainland China. Similar to Twitter in function and use, China’s Weibo has 50 million monthly active users in China. This project maps the differences in opinion espoused by the press and various reaction by netizens of social networks on both sides of the firewall.

5. Application program interface (API) is a set of routines, protocols, and tools for building software applications. An API specifies how software components should interact. Additionally, APIs are used when programming graphical user interface (GUI) components.

[3|4|5]
Objectives

The aim for the thesis is to create a visualization to illustrate the shape of the invisible wall that in turn shapes the influence of Internet censorship.

The core problem is a paradox that exists in the gap in people’s cognition which is caused by the absence of information. This paradox exists, and it is worthy to create a visualization about it so that more and more people can become aware of it. I looked to see if there are related projects and found there is no project that has visualized such kind of invisibility.

The form of the visualization is a node-link network diagram displayed on the web application. The design of the web application would enable users to share their opinions about the visualization with their connections on social platforms.

Target Audience

The target audience of this visualization are people who are interested in Internet censorship or people living inside of the wall and outside of the wall who are not aware of the invisible wall. Audiences would have a better understanding of Internet censorship and generate their opinions by exploring the network visualization and case study differences. The visualization would be web-based application so that users can share this project on their friend circle on social network pages. The audience would realize that information manipulation actually exists and close to their daily life.
Methodology

A comparison of public opinions from users inside and outside of the GFW was visualized by creating and properly using visual languages. The study followed a chronological research process, starting with a definition of the problem, data exploration, a literature review, and visualization design and develop. First, research on the history and purpose of censorship, including Chinese censorship history and Internet censorship history was conducted. During this process, the background research introduced basic knowledge, such as the history of the GFW, and general problems that influence of Internet censorship to social media and public opinion, which provided clear understanding of the topic for analysis. Defining the problem fully was necessary to develop the research before initiating data exploration and the literature review.

To avoid the pitfalls mentioned in Munzner’s research paper, Process and Pitfalls in Writing Information Visualization Research Papers, searching all related visualization projects and studies on the Internet was necessary. Previous researches had been conducted on the GFW, social media comparison, network visualization, and theories and case studies on online public opinion. (See the next chapter State of the Art) Overall, this step generated solid evidence and well-reasoned justification for the chosen study direction and visualization process for the project. The primary processes for visualization design included iterations of data discovery and user experience feedback. Design and development are necessary to move a project forward and require technological support to solve the research problem. These processes change and adapt during the iteration process.

Design for
User-centered
Data Visualization
Data exploration is an important step that specifies the problem and determines visualization design, which must fit the data features. To accurately interpret evidence, the shape of the data should be known so that visual representations match the data structure. Moreover, the research problem cannot be specified without data exploration. Necessary programming skills are required for data collecting, processing, storing, and web-front visualization development. Useful tools, like Tableau and R, generate quick visualization prototypes for data are useful for the data exploration. User personas and use cases provide more accurate visualizations. The ideal user audience should be specified, including cultural background and preferred devices for online browsing.

An excellent visualization design project matches information design to the research problem. Here, the users and implementation devices limit primary visual and technical possibilities, such as layout and interaction. The volume and velocity of data decided the statistical features and the conflict points. If an information design is visualization of a topic, conflict points in the data are what attracts and retains the audience. The variety and veracity of the data should also be considered to ensure the research is persuasive rather than neutral.
Background

Censorship History

The first dynasty of Imperial China, the Qin dynasty, was established by the first emperor Qin Shi Huang in 221 BC. Since then, centralized government has adopted information supervisory systems as tools to maintain the stability, solidity, and unification of Chinese society. To unify thoughts and maintain the authority of the central government, Qin Shi Huang adopted the policy of burning books and burying scholars to promote an official philosophy and intensify the power of his reign. His controversial dominant strategies of blocking other philosophies to promote an official ideology had profound effects on later rulers, who monitored opinions of political or religious governance. For instance, the Index of Prohibited Books was first published in 1559 before finally being abolished in 1966 by or the Roman Catholic Pope. It was a regulation of censorship and people who violated the official ideology were censored, investigated, or punished. With the advent of mass media, governments began to establish censorship department for publication and postal mails, especially during the first and second world wars in the twentieth century.

Figure 3


Internet Censorship

Currently, most developed countries use Internet censorship mechanisms, which were originally justified by governments to protect critical information, defend against cyber-attacks, prevent Internet crimes, and filter out harmful information. Political intentions soon became involved in censorship, which has resulted in political interference and social stability maintenance. Especially after the Arab Spring in 2010, China, Iran, and Egypt strengthened their censorship by adding more suspicious websites to their block lists and restricting more sensitive words in the surveillance of private messaging (see in Figure 3).

As the schematic diagram of a general communication systems\textsuperscript{10} shows in Figure 4, messages are transferred through transmitters distributed among the numerous node links of a network. Accordingly, online censorship could occur during any part of the telecommunication process. Regarding Internet blocking and filtering techniques, censorship occurs during the process of information passing between nodes, Internet users, and network links\textsuperscript{11}.

Censored content can include anything a user encounters daily on the Internet, such as websites, emails, P2P\textsuperscript{12} file sharing, video sharing, texting and messaging, VoIP\textsuperscript{13} (e.g., Skype), and social networks. Generally, censorship is not used to monitor everything that most ordinary users do because the larger the scale of users being censored, the greater the variety of information that is monitored, which requires more complex structures and advanced techniques\textsuperscript{14}. Each type of censorship requires different costs and accuracy depending on where it occurs during the information process.

Since first emerging between 1999 and 2002, three techniques (i.e., IP blocking/filtering, DNS hijacking, and content filtering) have been applied to Chinese censorship. The Chinese government has invested a huge amount of money and hired talented information technology (IT) professionals to the firewall project, which has the most sophisticated and complex firewall structure in the world. There are several different methods for filtering content, and while other totalitarian regimes have utilized one or more methods, it has been reported that only China exercises all of them.\textsuperscript{16}

---


\textsuperscript{12} P2P: Peer-to-peer file sharing, a distribution and sharing of digital media using peer-to-peer (P2P) networking technology.

\textsuperscript{13} VoIP: Voice over Internet Protocol, a category of hardware and software that enables people to use the Internet as the transmission medium for telephone calls by sending voice data in packets using IP.

\textsuperscript{14} IP: Internet Protocol

\textsuperscript{15} DNS: Domain Names System, maps internet domain names to the internet protocol (IP) network addresses they represent and enables websites to use names, rather than difficult-to-remember IP addresses.

History of Chinese Internet Censorship

Before Censorship

Censorship

Influence of Censorship


1980 | First e-mail sent from China
“Across the Great Wall, we can reach every corner in the world.”

1987 | First e-mail sent from China

1987 | First e-mail sent from China

1995 | Internet entered China
China Telecom started to prepare for building a national backbone network for Chinanet.

1995 | Internet entered China

1996 | China Golden Bridge Network (ChinaGBN)
The China Golden Bridge Network (ChinaGBN) opened a 256K dedicated circuit connected to the United States.

1996 | China Golden Bridge Network (ChinaGBN)

1998 | Initiation of the Golden Shield Project
The Ministry of Public Security took initial steps to control Internet use in 1997. In 1998, the China Democracy Party (CDP) was banned by government. The same year, the Golden Shield project was started.

2000 | Hackers and the Internet War
US and Chinese hackers engage in an internet war as the diplomatic row about the crashed US spy plane rumbles on. 56% growth rate.

2001 | Hackers and the Internet War
–26.5 million internet users

2005 | Golden Shield Project phase 1 was completed
The first part of the project lasted eight years and was completed in 2006.

2006 | Golden Shield Project phase 1 was completed

2008 | Golden Shield Project phase 2 was completed
The second part began in 2006 and ended in 2008.

2009 | China blocked Youtube, Facebook and Twitter

2010 | Google quitled mainland of China

2010 | Google quitled mainland of China

2014 | Instagram was blocked in mainland of China

2016 | Facebook Courts China With Censoring Software
China is a tempting market for Facebook, which has been banned there since 2009. But to get in, the social network may have to compromise on its mission.

2017 | Google is still in talks with Beijing over its plans to return to the mainland Chinese market, according to a senior Chinese lawmaker and former top official with knowledge of the negotiations.
Censorship on Social Media

Since the development and booming popularity of social media, an incredible amount of information has been created, allowing opinions and perspectives to be easily cited, shared, and discussed. Compared to public web portals and private email communication, social media has the advantage of engaging instant user activity and participation. Provocative opinions spread easily on social media, and opinion leaders are very influential within their network circles. Accordingly, governments have become more vigilant of this new type of communication.

Most of the information users create, discuss, and share can be censored during the data transformation process, including website views, community files and media sharing, personal emails, user texts and messages, and most recently, social media posts. However, due to the limitation of budget and operability, not everything on the Internet is under surveillance at all times. Censorship has different focuses in different domains. For example, some websites containing sensitive keywords related to terrorism might be blocked or surveilled, while an individual with a history of cybercrime might be censored at all times through his IP address or social media accounts.
State of the Art

Overview

This thesis provides a visualization of the influence of Internet censorship on public opinions. To do so, the following research questions were considered:

- Where can data about this subject be obtained?
- What is the scale of data that should be processed to provide accurate results?
- How can social network data be used to represent online public opinions?
- Are there any existing methods for data analysis?
- Is quantitative or qualitative research more appropriate for the purpose of this research in a social science field?
- What is the best way to visualize this comparative topic?
- Should the visualization be interactive or static?
- Should the visualization function as an analysis tool, an exploratory dashboard, or a data-storytelling presentation?

There are various prior visualization projects involving Internet censorship, social networks, and public opinions, which all use unique preferences and features to design visualizations, data selection processes, and exploratory or narrative functions. Each project has focused on a subject, study field, and target audience. This thesis used multiple criteria to cogitate these projects, which are listed below.
Study fields:
- Social networks.
- Public opinions
- Internet censorship
- Other fields of study.

Project property:
- Visualization.
- Research.

Visualization Design:
- Tool: Data tools are designed for data exploration and analysis without having a specific agenda or audience. Such tools can be used to find evidence through data exploration.
- Story-telling: Storytelling provides a narrative for audiences to get to know the point-of-view of a project.

Data attributes:
- Quantitative dataset
- Qualitative dataset

Data source:
- Social media API
- Journalism websites
- Dataset from other organizations
- Others

Information:
- Academic topics.
- Political information.
- Internet censorship related.
- Others.

Comparison:
- True.
- False.

This criteria matrix was used to review how relevant related works were to this thesis. There are research projects that provide multiple perspectives to show the power of government Internet censorship by demonstrating the availability of information, including website connection statuses and deleted content on social media (i.e., greatfirewallofchina.org, News Homepages’ Availability within China This Week, and Weiboscope). Some projects provide visualizations of social networks and online public opinions, which were referential literature for this thesis during visualization design (i.e., Political Grid Project, Twitter Sentiment Analysis App, Connected China, Weibo Visualization, and Unfiltered.news).
Greatfirewallofchina.org is a website providing tests of websites to see the real-time results that may be blocked in China. It aims to collaboratively build a community to visualize Internet censorship in an increasingly accurate way. From the website’s interface, a visitor types the URL of a website into a text box, and results are returned showing connections to different areas in China based on “Fail” or “Success” categorization. To do this, Greatfirewallofchina.org established several servers in China, and similar to a search engine, URLs are opened in servers located in China to return connection results.

Related Projects

Greatfirewallofchina.org
News Homepages’ Availability Within China This Week

This is an infographic project built by ProPublica, a non-governmental journalism organization based in New York City and established on November 17, 2014. It shows daily information about the availability of news homepages inside China. They chose 18 sites to test, either because of their international influence or their blocked history in China, which include BBC, CNN, WSJ, Reuters, Facebook, and top international news organizations and social media sites. The color-coded bars illustrate the censorship timeline for each website since November 17, 2014, and users can look up detailed information on a calendar view for each website. In addition to labelling websites as ‘blocked’ or not ‘blocked’, websites can also be ‘likely blocked’ or ‘likely not blocked’ based on existing ban frequencies. Contrasting colors, such as red for banned and green for approved websites, provide intuitive understanding of Chinese Internet censorship under the control of the GFW. The censorship data are obtained from GreatFile.org, which is a free service for testing website availability in China. The goal of this project is to continuously illuminate the status of news censorship in China in an easily understood way.

Weiboscope

Weiboscope is a Chinese social media data collection and visualization project, for which “the objective, among many, is to make censored Sina Weibo posts of a selected group of Chinese microbloggers publicly accessible.”

Weiboscope records blocked content for Weibo users with more than 1,000 followers or whose posts have frequently been censored since 2011. It provides real-time information about blocked tweets and keywords on Sina Weibo.

Weiboscope has developed its own data analytical and visualization tools, including censorship index time-trend analyses and term cloud analyses. These data tools have been used to undertake a variety of research projects, including analyzing Weibo censorship and its mechanisms, evaluating the impact of the real-name registration policy in China, determining social media and health issues in China, and topic modeling. Sina Weibo’s open API is used to access raw microblog data.

In terms of data visualization, under the ‘Censorship Index’ tab, a line chart shows a daily censorship index of blocked keywords and the number of blocked tweets containing these keywords. There are also word clouds illustrating the top one-hundred censored keywords on Weibo.
The Political Grid Project

This visualization project determines the influence of one of the most popular social media sites, Twitter, on political public opinion. Twitter data are collected from the Twitter API and users’ voting data. Users are asked to answer how much they agree with tweets and how important an opinion is to them. The results are shown in a grid scatterplot, in which the horizontal axis represents levels of agreement and the vertical axis represents degrees of care. Word clouds are also provided corresponding to candidate tweets to show a visual representation of general user impressions.

The diverse distribution of dots (i.e., user’s opinions) are clearly illustrated through the grid plots. Word clouds can effectively summarize different weighted perspectives. Although the amount of data is limited to a case study of the 2008 US presidential election, visualization tools can be used to represent public opinions using social media data.

Figure 13
Twitter Sentiment Visualization App

This website provides a web application that enables users to type in keywords and receive sentiment results in multiple visualizations, such as a scatter plot, heat map, word cloud, and timeline stack-bar charts. It also shows related recent tweets as data sources that users can browse. The authoring team devoted significant time to improving algorithms and sentiment analysis tools, which continue to work effectively for almost every topic on Twitter.
Weibo Visualization

This is an undergraduate data visualization project by Sylvia and Scott Chen, developed in 2012. It is an experimental visualization, based on a Sina Weibo dataset, which includes 1,425 users and 1,930 tweets on four topics. The purpose of the project was to "reveal the time-wise evolution of popular topics," especially the evolution of trends and popularity.

As one of the earliest data visualization projects focusing on Weibo, the complexity of the network and potential relationships between posts and users’ opinions on social media is well illustrated. Multiple overlapping charts on one dashboard with knotty may be confusing for both normal audiences and researchers; thus, it is important to develop improved interactions by filtering information based on user selection.
Connected China

Connected China is a collaborative project by visualization companies, Fathom and Thomson Reuters, developed in 2009. This web-based application is available on PCs and iPad and provides clear perspectives and insights about the dynamic Chinese political elite structure. The data are provided by Reuters, covering most of the leadership’s information about political life, relationships, and connections to each power group. It explains fundamental concepts of Chinese policy, previous and current leaderships, and related cultural celebrities and affairs.

The fluent interactivity and interface design enable project narrative and exploratory features. The design of the interactive infographics visualizes networks and hierarchies clearly. It is a great attempt to execute information storytelling of complex datasets by representing multiple related stories organized by a clear infographic skeleton.
On any single subject many ‘hear’ but few ‘listen’


Agenda Setting Theory

Agenda-setting theory describes the “ability of the news media to influence the salience of topics on the public agenda”18. It was first put forth by Max McCombs and Donald Shaw in their “Chapel Hill Study” of the American presidential election in 1968. They demonstrated a strong correlation coefficient between what 100 residents thought about the most important election issue and what the local and national news reported the most important issues to be.

In choosing and displaying news, editors, newsroom staff, and broadcasters play important roles in shaping political reality. Readers could know not only about a given issue, but also how much importance to attach to an issue based on the amount of information provided in news stories and the frequency they are mentioned. In reflecting on what candidates say during a campaign, the mass media may determine the important issues—that is, the media may set the agenda of the campaign.

Figure 17
Rogers and Dearing’s 1987 model of agenda-setting


Cultivation

When people are exposed to agitational information, they tend to agitate and think this world is more dangerous than it is.

“Some, normally the better educated and most politically interested (and those least likely to change political beliefs), actively seek information; but most seem to acquire it, if at all, without much effort.”

When journalists and media set the same agenda for their audience, the reader tends to believe that this agenda is the truth without hesitation. Over time the barrier of artificial information suddenly disappears, creating a divide between people’s ideologies.

Uses and Gratification

The audience does not always play a passive role in receiving information. The uses and gratifications theory (UTG) is an approach to understanding why and how people actively seek out specific media to satisfy specific needs. The driving questions of UGT are: Why do people use media and what do they use media for? The UGT discusses how users deliberately choose media that will satisfy given needs and allow one to enhance knowledge, relaxation, social interactions and companionship, diversion, or escape.

It assumed that audience members are not passive consumers of media. Rather, the audience has power over their media consumption and assumes an active role in interpreting and integrating media into their lives. Unlike other theoretical perspectives, the UGT holds that audiences are responsible for choosing media to meet their desires and needs to achieve gratification. This theory would then imply that the media compete against other information sources for viewers’ gratification.
Influence of Media to Public Opinion

The desires and thoughts of the most of people—or the collective opinion of a society on an issue or problem—is called public opinion. Before modern transportation developed, people used to live in a circle of acquaintances and did not travel great distances with frequency. News consisted of what people cared about the most and was primarily acquired through discussions with acquaintances. Limited by slow printing techniques, mass media did not exist to attract people’s attention and influence public opinions. The audience who read media consisted of a small group of highly educated people. Opinions were generated by them and spread to others through public speeches or discussions.

After the facilitation of modern transportation, it becomes easier for people to travel to distant places. Instead of local news, people began to care more about news from other parts of the world. Telecommunication techniques, such as the telegraph and television, accelerated the speed and volume of information transportation, allowing news media to publish more information. Advanced printing technologies and a larger educated population helped mass media grow, as more and more people began to get the news directly from such media.
Public Opinion on Social Media

With the development of Internet technology, it has become an inevitable that everything else has gone digital. In the digital era, increasing information is becoming easier to access and because of the leisure and entertainment properties of Internet, users spend more time immersed in the online world. The booming Internet companies such as Facebook, Twitter, and Weibo prove the successes of social media.

When people are surrounded by mass media, they take in information and opinions passively by reading articles or glancing at titles. People tend to be influenced by the opinions that journalism and media support, especially users who lack education and whose opinions fluctuate based on the media’s opinions. Based on the Annual Report on Development of New Media in China (2015) published by the Chinese Academy of Social Sciences, most users on Weibo are uneducated youth with low incomes; 55.18% of users are 10–21 years old, and more than 71.55% of users are high school students or younger.\(^{20}\)

---

Experiments Process

To design a visualization for Internet censorship, a methodology and series of experiments to obtain comparable datasets will be conducted.

First-round Experiment

Hypothesis

To show the influence of censorship, social networks will be used as data sources because they are good representations of public opinion and current trends. Social networks provide data about people’s opinions using comparable parameters, such as number of likes, forwards, and comments on a tweet and changes in trending topics and content of tweets. China has a different social network within the country than the rest of the world; thus, data about similar topics were collected from both Sina Weibo and Twitter to explore and determine gaps in people’s cognition using data visualization.

Sketches

Figure 22-28 shows the initial concept maps connecting concepts and ideas about Chinese Internet censorship and user behaviors on social networks. These concept maps provide possible aspects for data discovery and visualization design. Based on concept maps, primary sketches mainly focused on comparison of two social networks to determine possible methods of visualization.
In Figures 8 and 9, the concept maps reveal related concepts for the visualization project and their correlation. Based on these concept maps, the scope of content for visualization was limited to comparative information from social networks, which would provide updated information revealing people’s perspectives using quantitative features for the benefit of further analysis.
Data Processing

The first experiment was to obtain trending data from both Weibo and Twitter. Inspired by the Weibo Visualization and Twitter Sentiment Analysis projects, an API was used to obtain tweets and trends from social media. Compared to traditional journalism, users on social media participate in the process of information communication. However, obtaining desired trending data directly from social media was unpractical because it was difficult to find legal and free trending data for the past several years. China passed a cybersecurity law to define Internet management inside China, and with the audit standards increased, it has become very difficult to apply the Weibo API keys for the purpose of obtaining public data. Thus, the API key was attempted multiple times but no requests were approved for this research. For Twitter, the API for public usage was easily obtained but only allows developers to view data from the past week, which was too short of a timeframe for this experiment. Additionally, social media data amounts to a massive amount of information, which would take a long time to process, filter, and store.

To solve these issues, trending data were taken from two journalism websites using web-scraping tools. These websites contained news about hot topics on each social platform from 2013 until present. Trendistic is another news website about hot news on social media; thus, data scraped from Whatsonweibo and search results for “Twitter” on Trendistic were the samples used in the initial experiment. The content of the data was English, which was easier to process and analyze. The data were stored in a .csv file containing thousands of records for news on Sina Weibo and Twitter. The API from MonkeyLearn, an open machine-learning service provider, was used to process the text content and identify keywords and categories for the news. Then, d3.js was used to develop prototypes to explore the data.

Figure 29
Data process diagram
After obtaining the raw data, I used R to format the dataset and MonkeyLearn was utilized to determine the meaning of the data based on keyword extraction and topic classification. Then, the results were imported into D3 JavaScript to build exploration tools. In phase one of the process, bar charts showing all keyword rankings by time and relevance were generated. The keywords were extracted from the titles of news stories from both Weibo and Twitter by machine-learning tools. Each title had three to four related keywords, and colors represented the relevance of keywords to titles. However, the relationship and comparison of social media datasets were not shown.

Visualization Process
In phase two, donut charts provided categories for new titles on Weibo and Twitter. Each title was assigned to three to four related categories, and corresponding relevance was determined. There were common categories that were popular on both social media platforms, such as food, entertainment, and technology. The differences in trends were more difficult to observe; however, the connections between similar categories were not clear.

In the third phase and later prototypes, a node-link diagram was used to show the connection between news and categorization of topics. Color-coding differentiated news from each platform and by topic. The data were arranged through time to create a time-line visualization. The hypothesis is that the
diversity of topics on the two social networks would give the audiences a sense of what was happening on the inside of the GFW. Three groups of nodes and links ran from the top to bottom of the page: red nodes represented web news ranking through time sequence, blue nodes represented Twitter news, and green nodes represented labels and links to corresponding title nodes. It was quite challenging to visualize a clear-structured network using d3 force layout.

Results of the Analysis

For the next several weeks, the design and implementation of the visualization will be improved for all topics. Audiences can explore tweets by time or frequency. Because the dataset contains Chinese and English text, the visualization should use translation tools for better understanding by non-Chinese speaking audiences. Considering the exhibition, both of the experiments will be integrated into the web application so that the audience can view and share the visualization.

Second-round Experiment

The final visualization illustrated common features rather than differences for both sides of the GFW because differing public opinions might not be caused by Internet censorship but by cultural differences or political relationships between countries. The second-round experiment began by collecting data for specific topics to conduct a case study.
Hypothesis

The hypothesis was altered to state that the significant impact of censorship on people caused differences in news and user tweets on both social networks. Four topics were selected for this case study. Opinions are expressed based on a user's scope of knowledge and experience, which comes from education and information from the media; thus, it was not possible to find topics that eliminated cultural differences.

Non-political topics such as “ISIS” and “panda” were universally hated or liked, respectively, by most of people in the world, resulting in similar public opinions and different aspects of the content. Political topics, such as “Ai Weiwei” and “Dalai lama,” resulted in significant differences in the amount of available information on the social networks. Finally, social topics, such as “two child policy,” “feminism,” and “VPN (virtual private network),” also resulted in differing public opinions based on both content and sentiment.

Data Processing

Because the second-round experiment focused on comparing public opinions of specific topics, the data provided comparable information from the two social networks using the same topics, user attributes, and timeframes. For Twitter, public data were easily obtainable through the Twitter API for the prior week. For Weibo, public data were accessible using web-scraping tools for selected timeframes, although the network connection could be lost due to the GFW. The maintenance and storage of the data were also challenging, and a script was used to query data from Twitter and Weibo to save the data into a database.

The desired raw data were taken from both social media platforms and contained tweets in English and Chinese during February 2017. These tweets discussed the same topics, such as “panda,” “ISIS,” “Dalai lama,” “Ai Weiwei,” “feminism,” “two child policy,” and “VPN.” The dimensions of the data included users’ information, tweet content, time, topics, social media source, and information type. Information type could be the latest tweets posted by normal users or news tweets posted by news accounts on social media. Then, the text was processed into a content analysis in R.

Because most tweets on Weibo are written in Chinese, a word segmentation package for Chinese content was used to extract frequent words and classify tweets based on similarity in content for each topic. The English tweets were processed using the same package to obtain results with the same parameters.
Figure 36: screenshots of database in MongoDB

Figure 37: screenshots of processing code in R

Figure 38: screenshots of frequency words list
Visualization Process

Data Exploration

After data were trimmed, the resulting dataset containing frequency word lists for seven topics on Weibo and Twitter was explored. The visualization had to be clear and show the features of the data; thus, Tableau software was used to make prototypes and adjustments. In the figures below, bubble charts show the distribution of frequent words about the topic of feminism on both social media platforms. The bar chart figure illustrates the frequency ranking for the words. These simple visualizations depict attributes of data; however, it was necessary to enable the language translation function so that non-Chinese-speaking audiences were able to understand the Weibo data. Visualization should not only describe the data but also display common connections and differences between the two social media platforms.

Figure 37
First-round visualization

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Weibo data</th>
<th>Twitter data</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>200</td>
<td>300</td>
</tr>
<tr>
<td>500</td>
<td>600</td>
<td>700</td>
</tr>
<tr>
<td>1,000</td>
<td>1,100</td>
<td>1,200</td>
</tr>
<tr>
<td>1,500</td>
<td>1,600</td>
<td>1,700</td>
</tr>
<tr>
<td>2,000</td>
<td>2,100</td>
<td>2,200</td>
</tr>
<tr>
<td>2,500</td>
<td>2,600</td>
<td>2,700</td>
</tr>
<tr>
<td>3,000</td>
<td>3,100</td>
<td>3,200</td>
</tr>
<tr>
<td>3,500</td>
<td>3,600</td>
<td>3,700</td>
</tr>
<tr>
<td>4,000</td>
<td>4,100</td>
<td>4,200</td>
</tr>
<tr>
<td>4,500</td>
<td>4,600</td>
<td>4,700</td>
</tr>
<tr>
<td>5,000</td>
<td>5,100</td>
<td>5,200</td>
</tr>
<tr>
<td>5,500</td>
<td>5,600</td>
<td>5,700</td>
</tr>
<tr>
<td>6,000</td>
<td>6,100</td>
<td>6,200</td>
</tr>
<tr>
<td>6,500</td>
<td>6,600</td>
<td>6,700</td>
</tr>
<tr>
<td>7,000</td>
<td>7,100</td>
<td>7,200</td>
</tr>
<tr>
<td>7,500</td>
<td>7,600</td>
<td>7,700</td>
</tr>
<tr>
<td>8,000</td>
<td>8,100</td>
<td>8,200</td>
</tr>
<tr>
<td>8,500</td>
<td>8,600</td>
<td>8,700</td>
</tr>
<tr>
<td>9,000</td>
<td>9,100</td>
<td>9,200</td>
</tr>
<tr>
<td>9,500</td>
<td>9,600</td>
<td>9,700</td>
</tr>
<tr>
<td>10,000</td>
<td>10,100</td>
<td>10,200</td>
</tr>
</tbody>
</table>
First Prototype

The focus of the experiments was on using visual language to show differences in public opinions on both side of the GFW. The first prototype was a pair of opposite bar charts containing ranked-by-frequency words lists. For example, Ai Weiwei was one example used in the prototype, the results for which were illustrated by a static infographic based on two bar charts of frequent words about the same topic found on Weibo and Twitter. Color-coding specified the source and sentiment for the words. Red represented Weibo and blue represented Twitter, while purple lines linking the two sides of the chart and indicated common and frequent words, green represented positive words, and orange represented negative emotions. More linked lines indicated increased similarity in public opinions between the two social platforms. The number index showed the total amount of tweets about the topic.

The visualization revealed common public opinions on both social media platforms; however, it should also express differences. The ranking of frequency should be less prominent in the visualization and the number of words reduced to avoid distracting the audience. Additionally, it was not accurate to compare two ranking bar charts with different scales (axis x) because the base number of total tweets was significantly different.
Second Prototype

Based on the first prototype, improvements were made to a comparable visualization model. The second prototype used multiple components that took into account users' perception processes. It also used a web-based application so that users were able to share results on social media and access the visualization anywhere and at any time through an Internet connection.

First, a timeline graph introduced the history of Internet censorship in China. Then, an interactive visualization containing topic selection, tweet scale comparisons from each social media platform, and account types (from news or normal users' posts) provided bullet graphs displaying public opinions.
The scale visualization used rectangles to represent total tweets from each social media platform. The bullet graphs used edited word lists made by manual annotations to create clusters of words according to meanings. Words with similar meanings were shown together under the same categories. Red represented Weibo data, blue represented Twitter data, and the positions of dots were based on frequency scores, which were calculated by the quotient of sum-up frequencies and total number of tweets. Users could hover over dots and click on words on y axes to look up details.
These experiments developed methodologies for visualizing public opinions using comparable social media data. The differences in public opinions were strong when the distance between bullets was long, and public opinions were similar when distances were short. The scale of total tweets and percentages of news and user tweets revealed the popularity of topics, which could be manipulated by content censorship or the influence of censorship, especially for abnormal results, such as those obtained for the Dalai Lama and Ai Weiwei.
Future Work

Design and Development

For the next several weeks, the design and implementation of the visualization will be improved for all topics. Audiences can explore tweets by time or frequency. Because the dataset contains Chinese and English text, the visualization should use translation tools for better understanding by non-Chinese speaking audiences. Considering the exhibition, both of the experiments will be integrated into the web application so that the audience can view and share the visualization.
References


