Visualizing the Invisible: Censored Materials in China's Digital Space

Faye Ip, Sophia Lay

Introduction:

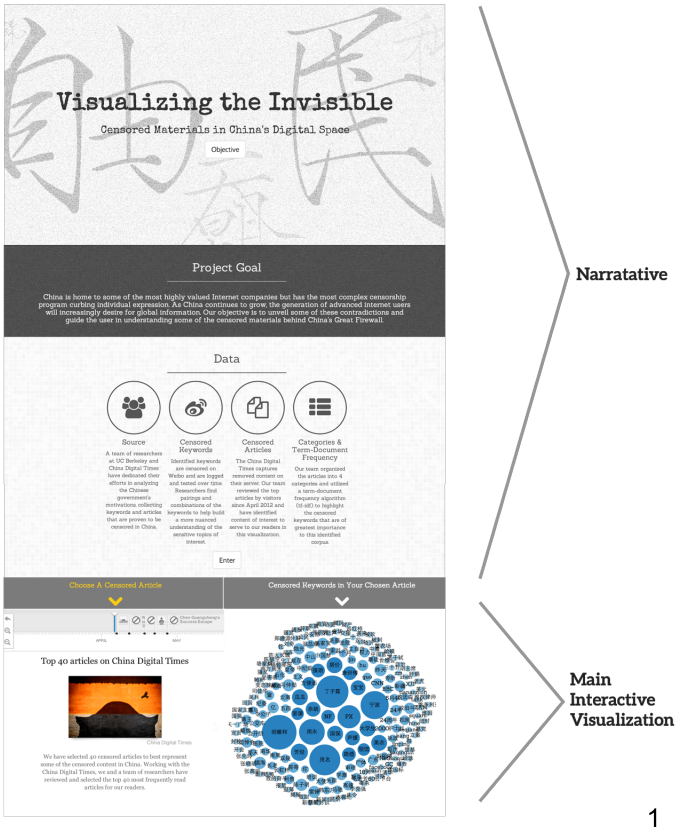
The goal of this visualization is to educate the general audience about internet censorship in China. The interface of the project is to guide users in understanding:

1. What is our project and problem statement about internet censorship in China.
2. Our approach in handling the dataset.
3. Important news events and articles that censored, and sensitive keywords that are blocked on Chinese search engines and social media.

Our intended audience is the general Western public, who are likely to be non-Chinese readers perhaps curious and interested in learning about this topic. While we are tackling a very complicated and difficult issue with layers of both technical and political complexities, we hope to help the reader discover some of the most sensitive news items of modern China and use our visualization on keywords to highlight how censored keywords relate to their contextual news events.

Discussion of related work

Our work dovetails with that of Professor Xiao, who collects censored materials leaked from China on his website chinadigitaltimes.net. A lot of scholarly work has been done on the topic of Chinese censorship, the large-scale efforts of the Chinese government in suppressing speech and stifling collective action, and the social and organizational forces that enable top-down directives from a centralized, authoritarian government onto an inherently decentralized, uncoordinated computer system which we call the Internet. As part of our



Visualizing the Invisible

Censored Materials in China's Digital Space

Objective

Project Goal

China is home to some of the most highly valued Internet companies but has the most complex censorship program curbing individual expression. As China continues to grow, the generation of advanced internet users will increasingly desire for global information. Our objective is to unveil some of these contradictions and guide the user in understanding some of the censored materials behind China's Great Firewall.

Data

Source: A team of researchers at UC Berkeley and China Digital Times have dedicated their efforts in analyzing the Chinese government's motivations, collecting keywords and articles that are proven to be censored in China.

Censored Keywords: Identified keywords are censored on Weibo and are logged and tested over time. Researchers find pairings and combinations of the keywords to help build a more nuanced understanding of the sensitive topics of interest.

Censored Articles: The China Digital Times captures removed content on their server. Our team reviewed the top articles by visitors since April 2012 and have identified content of interest to serve to our readers in this visualization.

Categories & Term-Document Frequency: Our team organized the articles into 4 categories and utilized a term-document frequency algorithm (tf-idf) to highlight the censored keywords that are of greatest importance to this identified corpus.

Enter

Choose A Censored Article

Censored Keywords in Your Chosen Article

Top 40 articles on China Digital Times

China Digital Times

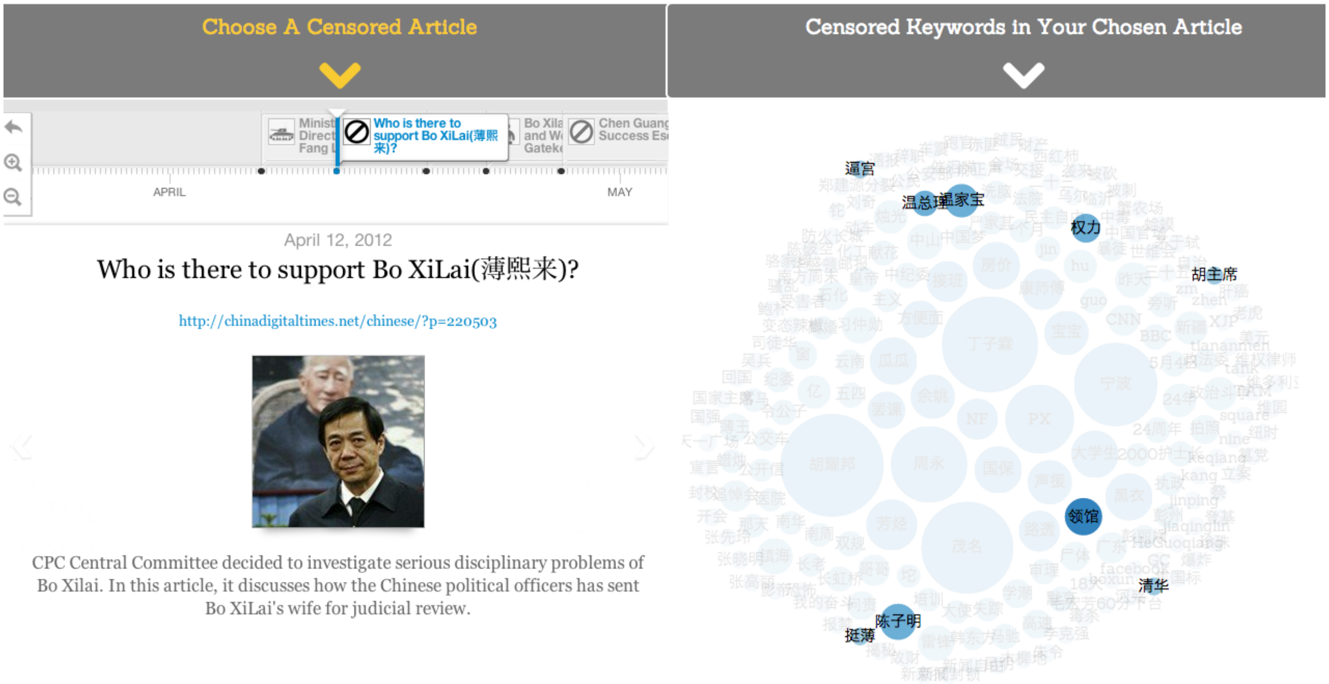
We have selected 40 censored articles to best represent some of the censored content in China. Working with the China Digital Times, we and a team of researchers have reviewed and selected the top 40 most frequently read articles for our readers.

preliminary research, we read the work of Professor Xiao and his research team, and also the work of other scholars, such as these research findings of a Harvard group which has done an extensive study on this topic.

Description of our visualization and the main narrative

We recognized that to shed light on the topic of Chinese censorship to our target users whom we assume have little background knowledge of China, a broad description of the scenario and an explanation of our data and methodology are necessary elements to accompany the main visualization. We adopted the design of a scroll webpage for users to go through the narrative parts first, allowing them to gain basic understanding and then providing them with the opportunity to explore the data in the interactive visualization part.

In the main interactive visualization part, our main design task is to link two different datasets together and make the relationship between the two obvious for users. We separated the article and keyword visual elements into two blocks, on the left and right sides of the screen.



Choose A Censored Article

Censored Keywords in Your Chosen Article

April 12, 2012

Who is there to support Bo XiLai(薄熙来)?

http://chinadigitaltimes.net/chinese/?p=220503

CPC Central Committee decided to investigate serious disciplinary problems of Bo Xilai. In this article, it discusses how the Chinese political officers has sent Bo XiLai's wife for judicial review.

Through our design process, we realized time is an important piece of information especially when the articles are highly related to news or social context. To display the articles with clear information of timing, we used timeline as the anchoring visual element. Timeline also serves as the filter for user to explore the further relation between censored articles and keywords. When user click on specific article, the below block will provide user with the following information:1) title of the article 2) url of the article 3) picture or video related to the topic of the article 4) summary of article translated into English. In addition to information of the article, the bubble chart on the right block will also grey out the keywords not contained in the article to highlight the keywords that appears in the selected article. Through this interaction, we achieve our initial goal to link two datasets.

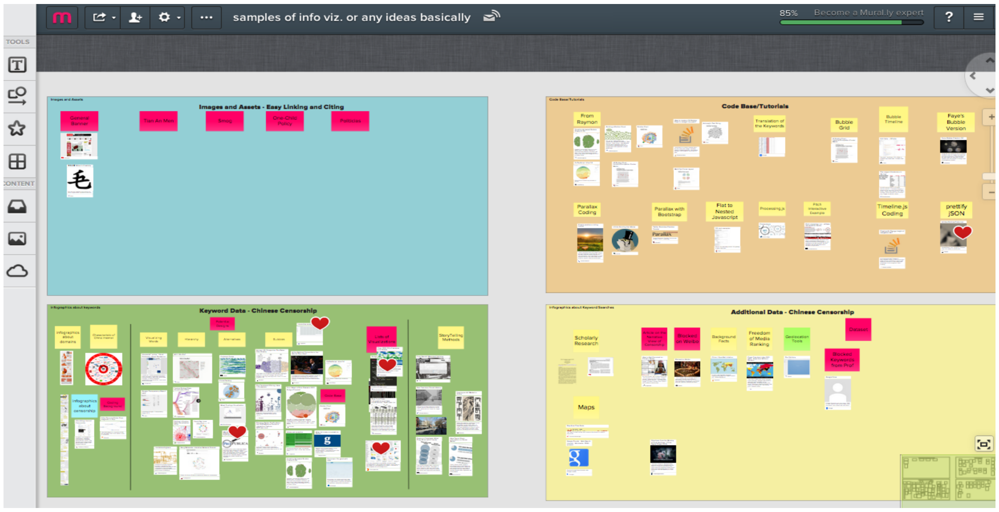
Data:

The data sets we worked with came from Professor Xiao's research team. They include: 1) a set of domain names blocked, and the methods by which they are blocked (e.g. DNS, IP, HTTP), 2) a list of sensitive keywords blocked from Chinese search engines and the Chinese equivalent of Twitter, 3) a matrix of articles vs keywords, displaying how many times each keyword appears in each article, and 4) a rank of all articles by number of visits on Professor Xiao's website. After many rounds of exploration, our team decided to focus our efforts on the keywords rather than the domain names because more data is available to support the visualization.

Tools:

Pre-production tools:

- **Mural.ly:** Our team used a collaborative poster-board tool to continuously share design ideas, code, research articles and background findings. We used this tool heavily throughout our pre-production stages to organize our sources and keep track of our ideas.

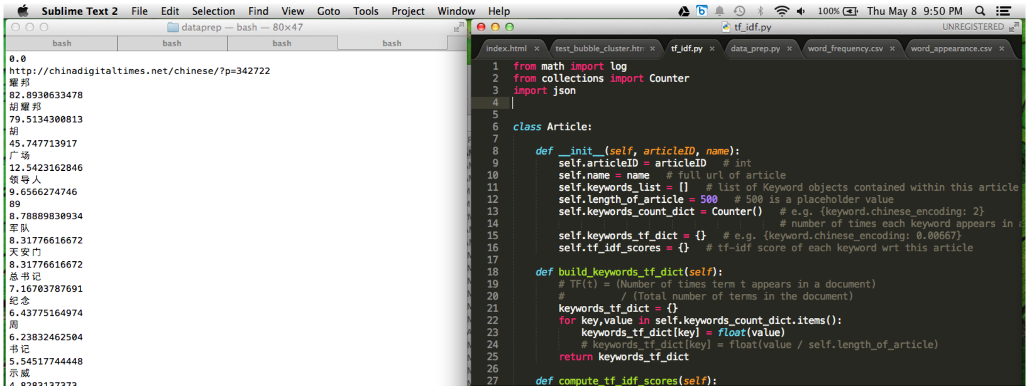

samples of info viz. or any ideas basically
85%
Become a Mural.ly expert
TOOLS
CONTENT
Images and Assets - Easy Linking and Citing
Code Base/Tutorials
Keyword Data - Chinese Censorship
Additional Data - Chinese Censorship
Bubble Grid
Timeline.js Coding
prettify JSON
Scholarly Research
Maps

- **Sketch:** During prototype stage, we used a graphic editing tool called Sketch to build out two visualization ideas and further gained users' feedback on these prototypes.

Production tools:

Upon reviewing some of the elements that our team was planning to build, we have defined our core elements to be an interactive timeline and a bubble chart that can provide insight about the censored keywords and their context to our user.

- **Timeline.js:** For the timeline, we reviewed some of the existing designs and timeline packages and landed upon Timeline.js. Its simplistic design made it very nice to use, but there were many aspects of it that didn't fit our intended design and our team had to work in tweaking many elements of it to make it fit to our needs.
- **Algorithm for ranking keywords**:



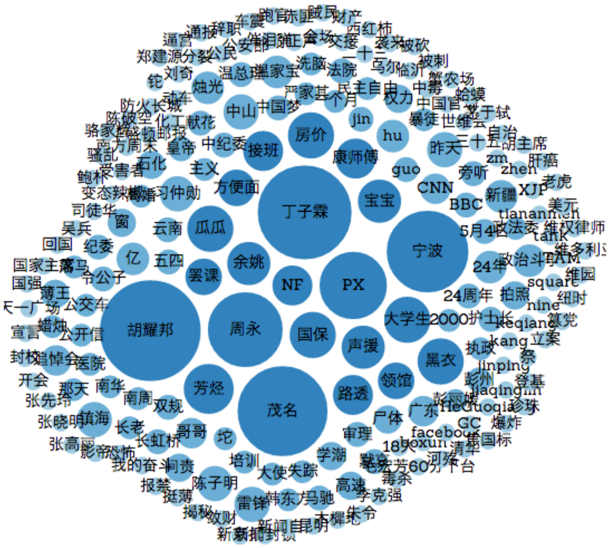
```
0.0
http://chinadigitaltimes.net/chinese/?p=342722
耀邦
82.8930633478
胡耀邦
79.5134300813
胡
45.747713917
广场
12.5423162846
领导人
9.6566274746
89
8.78889830934
军队
8.31776616672
天安门
8.31776616672
总书记
7.16703787691
纪念
6.43775164974
周
6.23832462504
书记
5.54517744448
示威
4.8283137373
```

```python
from math import log
from collections import Counter
import json


class Article:

    def __init__(self, articleID, name):
        self.articleID = articleID   # int
        self.name = name   # full url of article
        self.keywords_list = []   # list of Keyword objects contained within this article
        self.length_of_article = 500   # 500 is a placeholder value
        self.keywords_count_dict = Counter()   # e.g. {keyword.chinese_encoding: 2}
                                               # number of times each keyword appears in a
        self.keywords_tf_dict = {}   # e.g. {keyword.chinese_encoding: 0.00667}
        self.tf_idf_scores = {}   # tf-idf score of each keyword wrt this article

    def build_keywords_tf_dict(self):
        # TF(t) = (Number of times term t appears in a document)
        #         / (Total number of terms in the document)
        keywords_tf_dict = {}
        for key,value in self.keywords_count_dict.items():
            keywords_tf_dict[key] = float(value)
            # keywords_tf_dict[key] = float(value / self.length_of_article)
        return keywords_tf_dict

    def compute_tf_idf_scores(self):
```

Our team experimented with many different ways to rank the keywords in order to associate each keyword with a quantitative value that can be embodied by the size of each visual element representing the keyword.

- ○ **Frequency:** We first tried to quantify each keyword by its frequency, i.e. how many times the word appears in the collection of articles. But we found that the most common words, one that often do not have much meaning, tended to rise to the top of the rankings.
- ○ **Betweenness Centrality:** We then tried to model the keyword relationships as a graph, with each link representing whether or not two keywords co-appear in an article. We ran some centrality algorithms hoping to find the keywords that are most centrally clustered, but found instead that, because the graph is undirected, the betweenness centrality score of any particular node is not much different than the number of edges the node has, hence ranking the keywords once again by how common they are across the collection.
- ○ **Tf-idf:** Finally we experimented with tf-idf, with the hopes that the more common keywords in the collection would get demoted in ranking. After the first prototype of our tf-idf script, we tested the results with Chinese-reading users, who told us that the top-ranking multi-character keywords of an article do in fact serve as a meaningful proxy of the article's context. However, the single-character words still did not have much meaning, even if they had a high tf-idf score. Therefore, we further filtered out those words by increasing the idf threshold such that a word is included only if its idf score is above the threshold, regardless of its tf score. The tf-idf processing script is written in Python and is included with this submission.

- ● **Twitter Bootstrap:** We used Twitter Bootstrap to help us on layouting and CSS styling. Our team had envisioned a scrolling narrative feature that would bring the user through some of our thinking process in developing this visualization.
- ● **d3.js (Pack Layout):** Our team explored various bubble layouts for our visualization from various sources, including Mike Bostock's d3 repository, and experimented with



various bubble charts. We tried the force directed bubble, zoomable circle packing, cluster bubble layout and more. At the end, we came down to the bubble chart layout (which is a version of the circle packing layout with a flattened hierarchy) because a lot of the bubble layouts were too technically challenging and we found the bubble chart layout to also offer a level of simplisticity to allow for easy reading.

Steps to Accomplish Goals and Team Delegation

The process of development was split up roughly in the following stages. Each stage does not represent our meeting count. Our team meetings were mostly work-in meetings and we established a fluid working structure, where individuals would pitch in or contribute when needed. Team delegation is very collaborative and unified, but the rough distribution of work can be explained as follows:

Project Stage	Team Delegation
1) Project Background Discovery: We started by reviewing datasets from Professor Xiao, his team of researchers and students in his class. We discussed the data thoroughly to understand the source and methodology of acquiring that data.	Everyone
2) Data Exploration: Internally, our team defined our preliminary key goals and then divided the data sets to each teammate to explore. We then met to discuss our main key findings and started sketching on some preliminary ideas on how to visualize our findings from the data.	Domain dataset Keywords dataset
3) Initial Lo-Fi Sketching and Testing: We refined our sketches and then shared them amongst our team. We gave each other comments and then tested each design on two users. Based on what we learned from our users, we found that our users were not gaining the level of learning that we were expecting. Users were confused and had a hard time understanding keywords without the background context. Complex visualizations threw users into a process of extrapolation instead of gradual exploration and	Everyone sketched and took part in all usability tests.

learning. Our team deliberated and decided that we had to return back to our objective and redefine our narrative structure and storyline. It was a difficult stage as we knew that we had to redo some of these processes. We saw some novelty in trying to turn our lo-fi prototypes into hi-fi prototypes to see if there is a stronger storyline from our datasets.	
4) Hi-Fi Prototype: We decided to refine all 3 sketches to represent the data as accurate as possible by plugging in the data for hi-fi prototypes. However, we found that even within our dataset, the data alone did not tell a compelling story. We realized that the storyline we wanted to portray was in fact a living artery threading through political and historical events conveyed in the news articles; the keywords alone did not do it justice. We were all in agreement and returned to our objective stage to redo our narrative structure.	1: Progressive disclosure for Domain data 2: Keyword Node Graph 3: Keyword Matrix layout
5) Redefined Objective and Expert Review: Our team went back to do research, studied papers, consulted friends and reached out to Professor Xiao and Elisa Oreglia (an I School PhD student whose research focus is on China) for more insights on the topic. It was through our second round of research that really allowed us to drill deeper into the topic, ask the right questions and discover what is more important to show. Eventually, we refined our key points and visualization objective.	Everyone was present.
6) Storyboarding and User Scenario: After that, we each came up with a storyboard, wrote down our own objective and what we wanted our visualization to accomplish for our users. We all came with similar findings to educate the reader in learning some of the most sensitive	Everyone made a storyboard and returned to the meeting to discuss

items that are censored in China and aid the user in understanding how censored keywords can only be understood in its context. We also decided to filter down our data so that the visualization, in whatever design embodiment we choose, will not be inundated and cluttered to the point of overwhelming the user. We then structured our narrative.	
7) Design Concepts and Sketching: Upon defining what we want our visualization to accomplish, narrative and objective, we went on to explore certain visualization concepts that we wanted to use for our visualization. We sketched 3 designs, deliberated, then selected two and moved onto a mock-up stage.	Everyone sketched 1: Mock-up
Production Stage	
8) Data Manipulation and Keyword Categorization: With the vast amount of keywords, our team had to come up with a strategy of tying the keywords with its context. To get that type of data, we reached out to Professor Xiao's research team to find additional ways of adding quantifiable qualities to the keywords, such as list of articles on the blog and a count of the censored keywords in each article. Overall, there was a vast amount of data on the censored materials and keywords. In receiving the data, we divided the team: data exploration task and visual and content tasks. Upon receiving insights from the data exploration team, we are hoping that it can help guide the content team even more.	1: Data manipulation 2&3: Finding Code Sources, Reviewing Keywords and Categorization
9) Centrality Analysis + Layouting: The data set was immense and we soon found it difficult to find any insights. For that reason, we designed several methods of filtering the data. We looked at the data by word frequency, but it was not insightful. Afterwards, we decided to look at the data by betweenness centrality to see if it can help surface	1: Graph Analysis 2: in charge of the parallax scrolling and bubbles.

keywords that are interesting. Meanwhile, the other team focused on the other core visualization aspects (Timeline and Bubble Chart). Sophia was in charge and created our front page design that integrates the narrative background with the meat of the visualization.	1: timeline integration, banner design
10) Hi-Fi Prototype of Bubble Chart: Upon reviewing the centrality analysis, we still found some of the resulting keywords not as interesting as what we had imagined. For that reason, we decided to employ a qualitative method to help refine our data and guide our reader through selected content instead of having to find insights from the data.	1: Circle Packing 2: Cluster Bubble 3: HTML Templating and Timeline.js
11) Deep Qualitative Assessment + Content Review Having done centrality and frequency analysis, we knew we were having a lot of data problems because we didn't seem to be able to find interesting trends from the keywords. Our team met with Professor Xiao to discuss some of the ways to filter content qualitatively and the possible metrics that we can use. For that, he encouraged us to use the 'most frequently' visited articles metric to aid us in filtering articles. We discussed the validity of this metric and decided to use this method of filtering and observe the results. After that, we are hoping to continue our efforts in categorizing the content for display.	1: Circle Packing (Continue) 2: Cluster Bubble (Continue) 3: Article Selection + Content (Switched)
12) Bubble Chart, Styling and tf-idf Calculation Our team had immense technical difficulty trying to use d3.js to make bubble charts. We met with TAs and tried to have all hands working on the main visualization. We had trouble binding data to the nodes, linking nodes to the timeline and manipulating the timeline.js elements. For that, we kept questioning how to move forward and who to reach	1: Tf-idf Calculation 2: Bubble Chart and Styling 3: Content Insertion, Timeline and Styling

out to for help and advice. At the end, we decided to revisit what we want to accomplish and scale down appropriately to meet our ability. Our team came to recognize that our visualization doesn't need to be 'flashy' but must serve the main purpose of 'providing a broad contextual overview of the corpus' and also 'simple' and not overbearing on the timeline that we wanted to emphasize. For that, we decided to use the circle packing layout from d3. Meanwhile, we continued our main roles, while switching out Faye to work on the tf-idf calculation. At this point, we have solidified or main visualization and code layouts and then focused on reproducing what we need to accomplish.	
13) Tf-Idf Discovery, Article Translation and Styling Having received the results from Faye's tf-idf calculation, Sophia went ahead to review the results and see if the results were accurate and interesting. To our surprise, the top keywords were very interesting and provided insight into how the keywords are tied to the articles. For example, we found strange words to surface as important keywords to several articles; only later did we see that because a particular activist holds a very strange word and his/her username is a censored keyword. Meanwhile, with the 40 articles that we have qualitatively selected, Sophia went ahead to translate the articles in the timeline to provide some background context for the user.	1: Data Structuring and Organization, 2: Article Translation 3: Styling and Tf-Idf Keyword Reviewing
14) User Testing and Styling Having the visualization near completion, we did two rounds of usability testing to see if our visualization had any pressing problems for our users. We did find that users were very engaged with our timeline, but had a lot of trouble understanding our methodology and following the details of the dataset. The inherent assumptions behind	1: Styling 2&3: User Testing

the data falls into a very complicated subject matter we tried to work for a stronger narrative structure prior to the visualization for that reason.	
15) Styling and Integration of Core Elements While our team was working tirelessly together at all times, this phase was particularly crucial as we had to integrate all the elements together. At the beginning of this phase, we were still split between the tasks of data and styling, but close to the end, we had all hands on the bubble chart, working on jQuery to integrate the two pieces of the visualization together. jQuery was definitely a challenge for us, but having been able to complete this final task was a great victory for all of us.	1&2: Styling and Bubble Chart 3: Final edits on Data Manipulation. Everyone: Bubble Chart Integration with Timeline.

Usability Testing and Results:

Through our design process, we held two rounds of usability tests. First round was held during our initial brainstorming stage to test out the effectiveness of three different visualization based on lo-fi prototypes. Second round was held after we finished the design of whole web page to identify any key usability issues.

During our first usability test, we tested out three different visualizations to two target users. The most valuable insight we gained is that keywords don’t work on their own; users need context to understand the bigger picture of why certain words might be blocked or what the possible motivations are for the Chinese government to block them. This observation prompted us to rethink our data scope, and further decide to combine keywords and articles datasets together so as to give contexts to keywords by the narrative of articles.

From the second round of usability tests, we got a lot of positive feedback on how effective our design strategy of combining article and keyword enriches the context. Also, our design was praised by Professor Xiao for its ability to simplify the complicated dynamics of censorship.

However, we did find several usability issues that needs to be fixed in the future. First, the relationship between timeline and bubble charts need more visual clues. Second, the meaning of the radius of the bubble needs explanation. Third, users are highly interested to know the most prominent keywords in an overall scope instead of single-article scope. Last, the users are intrigued to learn the further information of the keyword, including translation of the keyword and the contexts of the keyword being used.

Conclusion:

This project attempts to introduce a Western audience to some of the complexities and contradictions of China's digital censorship efforts. The core visualization of the project bridges the relationship between sensitive keywords, blocked on Chinese search engines and social media outlets, and censored news articles, deleted or blocked at the behest of the Chinese government. This data is leaked by sources inside China without whose efforts these materials would remain invisible to the world, in an age where the Internet has already enabled unfettered access to information for a large majority of its users. Our overall project goal, therefore, is to offer a glimpse of what lurks beneath visibility, and to hint at what the possible motivations are for the government to keep certain things out of its netizens' reach.

Due to time constraints and technical limitations, not all of our design ideas were realized at this stage. Possibilities for future improvements include: using a navigation bar to divide the articles by important topics (e.g. "Tiananmen", "Corruption", "Political Leaders", etc), using colors to differentiate keyword nodes by topics, repainting a new svg with each topic click to refresh the bubble sizes, offering an English translation for each keyword, and offering a pop-up list for each bubble to link to all articles containing that particular keyword.

Appendix:

1. Link to Demo: http://people.ischool.berkeley.edu/~sophia.lay/infoviz_chinese_keywords/

2. Github Repository: https://github.com/fayeip/infoviz_chinese_keywords