Online Appendix to "Microblog Dissent and Censorship During the 2012 Bo Xilai Scandal"

Developing the Sentiment Categories

In contrast to similar recent projects (Cairns and Carlson 2016; Plantan and Cairns 2017) in which I developed sentiment categories and manually coded posts alongside a co-author and a research assistant (for a a three-person coding team), for this project I coded solo due to a lack of available resources. However, these two prior projects provided me with valuable practice in reading and coding *Weibo* posts, and increased my confidence in doing so alone this time.

I began the exercise by using structural topic modeling, or STM (Roberts *et Al.* 2014) to identify latent scandal-related topics in the *Weibo* data and key words/phrases associated with each topic.[1] I did this by first searching the WeiboScope corpus for the simple keywords "Bo Xilai" and "Wang Lijun" and then identifying peak dates for this keyword's incidence. I took all dates with keyword counts more than two standard deviations above the year-long mean, for a total of 11 days that would later correspond to the peaks of Phases I-III. I then separated out the text from these dates and used the Txtorg program (Lucas *et Al.* 2014) to create a Term-Document Matrix, or TDM.[2] Next, I input the TDM into the above authors' topic model, implemented in the R language.[3] After model estimation, I used STM's LabelTopics function to report lists of the most frequently associated keywords with each topic.

After looking at this algorithm-generated keyword list, I then drew and read through several random samples from each of the four time periods (February 8-12, March 14-16, April 11 and September 28-29) to see which keywords were strongly associated with what I judged to be topically relevant content, and which were just noise from the automated procedure. I whittled down the list of keywords to the final ones in the chapter, then based on these words and holistic post reading, formally defined the sentiment categories. Next, I drew a sample of 100 'practice' posts taken evenly from across the four date ranges. After going through this coding exercise, I refined the scheme and reduced the number of categories. Finally, I drew a sample of 1000 posts total (250 from each date range) and proceeded to score these according to the category definitions. These posts then served both to directly estimate category proportions for those dates, and as input into *ReadMe*.

[1] STM is a topic-modeling algorithm based on a family of unsupervised machine learning models called Latent Dirichlet Allocation, or LDA.

[2] A term-document matrix is a mathematical summary of the frequency of terms appearing in a text corpus, and is used as input into many natural language processing algorithms.

[3] The key researcher-chosen parameter in a structural topic model is the number of topics K. There is no "right" number of topics, but picking a nonsensically high or low number may lead to confusing or poorly interpretable results. After some experimentation, I settled on $K = 10$ topics. Later on, when I had switched from computerized topic modeling to manually reading the posts, I trimmed the number of sentiment categories ("topics") down to the five presented in this paper.

Correcting for Bias in Censorship Rate Estimation

One of the major difficulties in using Chinese social media data is how to deal with the bias induced by state censorship, since researchers attempting to "harvest" such data are only able to observe the blog posts they are able to download faster than censors can delete these posts. However, as long as researchers are able to capture a fraction of all censored posts, it may be possible to estimate the true censorship rate. First, assume that out of the sample of around 43,000 Weibo bloggers, some fraction decide to write a post in response to some event.[4] Also assume that individuals who choose to write a post do so immediately following the event.[5] What I want to know is how many of these posts will survive (not be censored) long enough to appear in the WeiboScope data. This information is necessary to calculate my primary quantity of interest – the true censorship rate in Equation 1:

$$R_{true} = \frac{C_{obs} + C_{hid}}{C_{obs} + C_{hid} + P} \tag{1}$$

Where R_{true} is the true rate, expressed as the proportion of censored over total posts, P is posts that are never censored (all of which appear in the dataset), C_{obs} is the number of posts marked as "censored" in the dataset, and C_{hid} is those posts that get censored, but do not appear in the dataset because they are deleted sooner than the Hong Kong team can download the *Weibo* user timelines that contain them. The WeiboScope data scraping process, as described in Fu, Chan and Chau (2013), involved periodically returning to the pages of the 43,000 users, downloading a copy of the timeline each time. If a post got deleted between crawls (i.e. after the team's program had crawled a page during a particular iteration, but before the next one), then the researchers could compare the new record to the old one, identify the post that had disappeared in the interim, and mark it as censored. However, due to limits set by Sina.com, the team could only crawl most of these pages (38,000 out of the 43,000, who constituted the "Verified" user group) once every 24 hours. Given a uniform distribution of sensitive post-inducing events (i.e., that they were equally likely to occur over a given 24-hour period), the average time between when a post would go up and when that Verified user's page would be crawled, would be 12 hours. Since Zhu et al. (2013) find that most censorship occurs within an hour or so of the post time, most censored posts from the Verified users were unlikely to make it into the dataset. Thus, the dataset is truncated, and R_{true} will be biased. What I have is the observed rate, R_{obs}, in Equation 2:

[4] The fraction that decides to write versus not write a post in response to breaking political news does not matter for modeling the data-generating process, and I do not consider it further because I only care about generalizing my findings to those individuals who do post – I do not seek to explain "participation" in the dataset.

[5] While this simplifies reality, the findings of the exercise here generalize easily to cases where individuals choose different durations after an event at which to write a first post, provided that posts occurring later on follow the same censorship distribution over time as their immediate counterparts.

$$R_{obs} = \frac{C_{obs}}{C_{obs} + P} \tag{2}$$

Since C_{hid} is missing, $R_{obs} < R_{true}$, i.e. the observed censorship rate is biased downward. But how much so? The observed year-long average rate for the topics in the author's other Weibo data work (Cairns 2017) is between 12% and 17%, an oddly low figure given that other studies (King, Pan and Roberts 2013) have measured the true rate during sensitive events to be closer to 60% and I have no reason to think that the present topic is any exception. To calculate the true rate, I need to know the true number of posts censored, N_{true}, which is related to N_{obs}, the number of censored posts that I actually observe, via some equation that models the speed with which censors remove posts during sensitive episodes.

Since I do not know the true equation, I need to look for an empirical example that provides a good approximation. The best available so far is the finding by Zhu *et Al.* (2013), who note that "nearly 90% of deletion events happen within the first 24 hours" (p. 1). Conveniently, this time window is the same as that of the unbiased portion of the data: 100 percent of posts will be observed, and correctly identified as censored or not, if they survive 24 hours or more. Since Zhu *et Al.* found that 90 percent of censorship occurs before 24 hours, 10 percent must occur after, sometimes days or weeks later. Since I observe this 10 percent, and critically, assuming that distribution of censorship over time is the same in my data as in that of Zhu *et Al.*, the ratio of what I observe to what gets missed must be 1:9, e.g. $C_{hid} = 9C_{obs}$. This suggests that multiplying C_{obs} by a factor of 10 will get me close to the true rate. Plugging this into Equation 1 gives Equation 3:

$$R^{*}_{true} = \frac{C_{obs} + 9C_{obs}}{C_{obs} + 9C_{obs} + P} \tag{3}$$

Applying Equation 3 to the censored post data gives Table 1 below:

Table 1: Observed Versus True Censorship Rates For Peak Discussion Dates: Bo Xilai Scandal (90%/24 hrs)

Date	Posts ($C_{obs} + P$)	Observed rate (R_{obs})	True rate (R^*_{true})
2/8	1745	.04	.28
2/9	3528	.03	.22
2/10	3754	.03	.23
2/11	5765	.01	.11
2/12	1462	.04	.30
3/14	1203	.11	.56
3/15	4338	.10	.52
3/16	1441	.09	.49
4/11	2385	.06	.40
9/28	1235	.18	.69
9/29	1114	.19	.70
All Phases (avg)	394	.17	.61

Given that I am applying another Weibo study's findings to a different dataset, the question might arise, given that my data consist of journalists, dissidents, and Verified users with more than 10,000 followers – all sensitive groups in censors' eyes – whether 90 percent within 24 hours is too slow a rate for the sample. Zhu *et al.* and King, Pan and Roberts both find that some small fraction of ultimately censored posts typically linger for days after an incident – the question here is how much. My main empirical concern in this paper is under-, not over-estimating the censorship rate. If I assume that the true number is 95 percent within 24 hours, i.e. $C_{hid} = 19C_{obs}$ then plugging these numbers into Equation 2 yields Table 2:[6]

[6] Given that the assumption of 90% post deletion within 24 hours is already very pessimistic, I believe 95% represents an absolute worst-case scenario. 90% within 24 hours would only be true if the entire year of 2012 were constantly filled with sensitive pollution-related online outbursts – it is unlikely that the in-house censors Sina employs to delete posts devote the resources and attention necessary to achieve such a fast deletion rate for non-critical events, although Appendix B.1 does explore this further. This is why I think my adjusted measure of censorship probably overestimates the true rate for much of the year. However, since for statistical purposes I am primarily concerned with censorship fluctuations rather than the level, the specific censorship adjustment I choose should have little impact on my results.

Table 2: Observed Versus True Censorship Rates For Peak Discussion Dates: Bo Xilai Scandal (95%/24 hrs)

Date	Posts ($C_{obs} + P$)	Observed rate (R_{obs})	True rate (R^{*}_{true})
2/8	1745	.04	.43
2/9	3528	.03	.37
2/10	3754	.03	.37
2/11	5765	.01	.20
2/12	1462	.04	.47
3/14	1203	.11	.71
3/15	4338	.10	.68
3/16	1441	.09	.66
4/11	2385	.06	.57
9/28	1235	.18	.82
9/29	1114	.19	.82
All Phases (avg)	394	.17	.73

The rates above are higher than the previous estimates. However, the estimated mean censorship rate for February 8-12 (Phase I) still hovers around 40% and goes as low as 20%, a rate far less than later dates.