**Online Supplementary Materials: Methodological Appendix**

Our novel dataset contains 2,677,072,844 Weibo tweets posted by 170,105,271 users during the year 2013.

*Identifying opinion leaders*

First, we picked the top 5,000 users in our dataset according to follower count. We then complemented this measure by using the Chinese Opinion Leader Ranking released by New Media. This ranking uses a Micro-blog Communication Index (BCI) to measure both user productivity and popularity. The BCI index is formulated as:

BCI = (20% \* W1 + 80% \* W2) \* 160,

where W1 is a measure of user productivity based on the number of original tweets the user produces:

W1 = 30% \* log (tweet count + 1) + 70% \* log (original tweet count + 1);

and W2 is a measure of user popularity based on the feedback the user receives:

W2 = 20% \* log (retweet count +1) + 20%\*log (comment count + 1) + 25% \* log (original retweet count + 1) + 25% \* log (original comment count + 1) + 10% \* log (like count + 1)

We combined the top 100 users in ten issues of this ranking and retained 311 non-repetitive names, most of which were already included in our top 5,000 list. However, most users in this long list (many are celebrities and brands) do not contribute to public discussions of political and social issues. Therefore, we applied automated text analysis to look for people keen on politics. First, we selected 20 widely-acknowledged opinion leaders who often talk about politics,[[1]](#endnote-1) used *all* their tweets in 2013 as a document, applied TF-IDF statistic[[2]](#endnote-2) to identify the top 1,000 keywords of their speeches, and finally selected 93 politically related terms from these keywords (see Appendix Table 1.)

Second, we matched all tweets of users in the long list[[3]](#endnote-3) with the 93 political words. For each user, we also calculated the ratio of tweets containing political words to a user’s total tweet count. Then, we selected users who had at least 15 tweets containing political words or whose political tweet ratio was higher than 5 per cent. This gave us a total of 204 individual opinion leaders who were, in general, concerned about political and social issues.[[4]](#endnote-4) They posted an average of 443 tweets in 2013, whereas average users only posted 25 tweets. Moreover, opinion leaders had a mean of 2,272,875 followers, whereas average users only had 31.7 followers.

*Data collection*

This study primarily focuses on opinion leaders’ posts on the US, Japan, Taiwan and the political system (*tizhi* 体制) in China. We collected data by matching tweets with keywords associated with the above topics. However, instead of simply determining keywords by local knowledge, we used a fuzzy matching algorithm to automatically expand the pre-defined keywords to similar words used by Weibo users. For instance, if only the word “美国” (America) was used to match tweets associated with discussions about the US, tweets using the alternative term, “美帝”(Imperial America), which is frequently seen as a humorous term referring to the US on the Chinese internet, would be missed. The fuzzy matching algorithm powered by a Word2Vec model[[5]](#endnote-5) can solve this problem by identifying similar words people use in this discursive space. We finally selected 13 keywords as suggested by the algorithm (see Table 1 in the main article.)

Again, we matched tweets from 204 opinion leaders with the 13 keywords, and we only selected opinion leaders with at least 5 tweets, or 3 per cent of tweets, containing these keywords. As a result, we obtained 6,087 tweets produced by 146 opinion leaders with regard to five nationalism-related topics. Note that we dropped Taiwanese opinion leaders (for example, Kai-fu Lee @李开复，Qiu Yi @邱毅台湾, etc.) from our data because their identity may bring complications to the discussion of “Chinese nationalism.” Detailed demographics of the 146 opinion leaders are shown in Appendix Table 2. Opinion leaders’ Weibo behaviour is summarized in Appendix Table 3.

*Hand-coding*

Owing to the insufficiency of computer algorithms when processing Chinese short texts, especially when a lot of jargon, banter, puns and coded language is involved (which is often the case when people talk about politics on Weibo), we relied on human coders to analyse texts. Our hand-coded data could be used to train better sentiment-analysis models for future studies.

We hand-coded 6,087 Weibo tweets using a five-level scale ranging from -2 to 2. Detailed coding scheme and examples are summarized in Appendix Table 4. We also asked a second coder to code a random sample consisting of 1, 243 tweets from 30 random opinion leaders. Inter-coder reliability statistics are given in Appendix Table 5.

*The issue of censorship*

To address the concern that the less popular status of nationalists may be owed to state censorship, we analysed another dataset consisting of censored Weibo tweets collected in 2012 by the Weiboscope project (see <http://weiboscope.jmsc.hku.hk/datazip/>). We have assumed that the environment on Weibo in 2012 was not dramatically different from that in 2013.[[6]](#endnote-6) This project collected the timelines of 350, 000 influential Weibo users – defined as users with more than 1,000 followers – including tweets which were censored later.

We sampled 921 (approximately 1 per cent) out of the total 86,083 censored tweets in the dataset. Specifically, we drew a 1 per cent random sample (Numpy random seed = 300) from each of the 12 months in 2012. For months when the 1 per cent sample included less than 35 tweets, we instead drew 35 random tweets to ensure statistical validity. The distribution of sample counts is described in Appendix Table 6.

We then hand-coded the censored sample into five categories. The coding scheme is illustrated in Appendix Table 8. The hand-coding was cross-checked by a second coder. Inter-coder reliability statistics are presented in Appendix Table 7. The results show that, out of 921 tweets, only 12 expressed nationalist sentiments, while as few as four vented anti-regime nationalist sentiment. By contrast, there were as many as 340 anti-regime non-nationalist tweets. Interestingly, many anti-regime tweets were attacking nationalist views. We then estimated the proportion of each category in the total population at alpha = 0.01 level. Results are shown in Appendix Table 9.

To further address the concern that censorship of nationalists may be disproportionately intense during nationalist movements, we also hand-coded all 551 censored tweets between 11 and 21 August 2012, a period when the Diaoyu dispute provoked large-scale anti-Japan demonstrations in China. Unsurprisingly, the proportion of nationalist tweets increased, yet not as high as expected – censored nationalist tweets accounted for 8.87 per cent during this period, 3.62 per cent of which were anti-regime nationalist tweets. By contrast, 27.71 per cent of tweets were anti-regime commentaries *without* nationalistic reasons.

Although the censored non-nationalist tweets largely outnumbered nationalist tweets, this finding is, by and large, consistent with King, Pan and Roberts (2013; 2014), who argue that the state primarily censors tweets associated with collective action potential. In our sample, about 36 per cent of censored tweets were anti-regime but *not* nationalist, while another 31 per cent were neither anti-regime nor nationalist. These two categories made up the vast majority of the censored data, and they mainly included tweets about emergent natural or man-made disasters that bore collective action potential. The only difference is that the former clearly criticized the government or officials for these incidents, whereas the latter simply disseminated information regarding these incidents.

All in all, analysis on the censored tweets shows that nationalists were not heavily censored. Therefore, the marginal status of nationalists on Weibo was not likely to be owed to state censorship.

Appendix Table 1: **Key Words Used to Identify Opinion Leaders**

|  |  |  |  |
| --- | --- | --- | --- |
| **Chinese**  | **English** | **Chinese** | **English** |
| 三中全会 | Third Plenary Session  | 邓小平 | Deng Xiaoping |
| 军事 | military  | 何兵 | He Bing |
| 军队 | Army | 王立军 | Wang Lijun |
| 美国 | United States | 民主 | democracy |
| 美帝 | Imperial America | 宪政 | constitutionalism |
| 日本 | Japan | 社会主义 | socialism |
| 朝鲜 | Korea | 法治 | the rule of law |
| 香港 | Hong Kong | 爱国 | patriotic |
| 钓鱼岛 | Diaoyu Islands | 普世 | universal values |
| 台湾 | Taiwan | 人权 | human rights |
| 叙利亚 | Syria | 言论自由 | freedom of speech |
| 埃及 | Egypt | 禁言 | banned post |
| 中日 | Sino-Japan | 敏感 | sensitive |
| 苏联 | Soviet Union | 福利 | welfare |
| 习近平 | Xi Jinping | 民生 | people’s livelihood |
| 毛泽东 | Mao Zedong | 文革 | Cultural Revolution |
| 毛主席 | Chairman Mao | 维权 | rights protection |
| 周永康 | Zhou Yongkang | 上访 | petition |
| 金正恩 | Kim Jong-un | 司法独立 | judicial independence |
| 袁裕来 | Yuan Yulai | 城管 | law-enforcement officer |
| 马英九 | Ma Ying-jeou | 腐败 | corruption |
| 薄熙来 | Bo Xilai | 污染 | pollution |
| 柴静 | Chai Jing | 环保 | environmental protection |
| 反腐 | anti-corruption | 体制 | political system |
| 谣言 | rumour | 言论 | speech |
| 贪官 | corrupt officials | 大国 | great country |
| 强拆 | demolitions | 运动 | social movement  |
| 劳教 | re-education through labour | 民意 | public opinion |
| 删帖 | censorship | 意识形态 | ideology |
| 食品安全 | food safety | 执政 | ruling (party/ruler) |
| 医改 | medical reform | 政权 | regime |
| 贪污 | corruption | 货币 | currency |
| 公知 | public intellectuals | 经济 | economy |
| 五毛 | Fifty-cents army | 城镇化 | urbanization |
| 公民 | citizen | 夏俊峰 | Xia Junfeng |
| 国人 | compatriot | 薛蛮子 | Xue Manzi |
| 二代 | second generation | 吴虹飞 | Wu Hongfei |
| 汉奸 | traitor | 秦火火 | Qin Huohuo |
| 敌人 | enemy | 立二拆四 | Li’er Chaisi |
| 富人 | the rich | 张雪忠 | Zhang Xuezhong |
| 精英 | Elite | 夏业良 | Xia Yeliang |
| 左派 | left wing | 许志永 | Xu Zhiyong |
| 宪法 | constitution | 陈永洲 | Chen Yongzhou |
| 冤案 | injustice | 王功权 | Wang Gongquan |
| 立案 | file a case | 斯诺登 | Snowdon |
| 政治 | politics | 曼德拉 | Mandela |
|   |   | 服贸 | Taiwan–China service trade agreements |

*Note:*

The 93 terms are key political words (identified by computer algorithm) that frequently appeared in well-known opinion leaders’ posts in 2013. Thus, a number of keywords are associated with significant incidents in 2013, e.g. Snowdon (NSA PRISM), Bo Xilai (coup scandal), Chai Jing (*Under the Dome*) and the like. Note that some words (e.g. 美国/United States) may bring noises into data as they are also high-frequency words in non-political contexts. To adjust for this, we asked a human coder to go over the results and remove irrelevant tweets. Table 3 reports the adjusted results.

Appendix Table 2: **Descriptive Statistics: The Demographics of 146 Opinion Leaders**

|  |  |  |
| --- | --- | --- |
|   | Frequency | Per cent |
| *Sex* |  |  |
|  Male | 133 | 91.09 |
|  Female | 11 | 7.53 |
|  Unspecified | 2 | 1.37 |
| *Verification* |  |  |
|  Yes | 138 | 94.52 |
|  No | 8 | 5.48 |
| *Region* |  |  |
|  Beijing | 84 | 57.53 |
|  Shanghai | 10 | 6.85 |
|  Guangdong | 10 | 6.85 |
|  Zhejiang\Jiangsu | 5 | 3.42 |
|  Other inland cities | 14 | 9.59 |
|  Overseas/Hong Kong | 12 | 8.22 |
|  Others | 11 | 7.53 |
| *Industry* |  |  |
|  Academia | 33 | 22.6 |
|  Business | 20 | 13.7 |
|  Professionals | 52 | 35.62 |
|  Grassroots | 35 | 23.97 |
|  Government | 6 | 4.11 |
| Total | 146 | 100 |

Appendix Table 3: **Descriptive Statistics: Weibo Behaviour of 146 Opinion Leaders**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Mean** | **SD** | **Min** | **Max** |
| Follower count | 2,272,875 | 4,904,651 | 24,799 | 3.45e+07 |
| Following count | 1,385.25 | 895.45 | 0 | 3,685 |
| Friends count | 1,143.53 | 791.16 | 0 | 3,623 |
| 2013 post count | 443.21 | 434.70 | 30 | 3,092 |
| No. of nationalism posts | 42.71 | 56.15 | 3 | 372 |

*Note:*

Following count is the number of users that a given opinion leader follows. Friends count is the number of followers who are also followed by the opinion leader (mutually following). 2013 post count is the number of Weibo tweets posted by opinion leaders in our 2013 data. Nationalism posts are Weibo tweets that contain nationalism-related keywords.

Appendix Table 4: **Illustrations of Coding Rules**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|   | **US** | **Japan** | **Taiwan** | **Diaoyu Islands** | **Regime** |
|   | *whether one recognizes and favours American culture, values, and political system* |  *whether one emphasizes past wrong-doings or co-development in the future* | *whether one favours Taiwan culture/political system or feels reunification is justified* | *whether one calls for taking the Diaoyu Islands back even with force* | *whether one recognizes and favours the political system and the ruling party in China* |
| -2 | American democracy is hypocritical and America is a unipolar enforcer of its own interests on others | We will never forgive the sins of Japan. Your current accomplishments have been built on China’s blood and tears. | Taiwan’s democracy is such a joke. Look at the worn infrastructure and broken buildings. If I were Taiwanese, I’d like to move to the mainland. | Defend Diaoyu! Diaoyu belongs to China! | If the empire is dismantled, people will lose nothing, but will be freed.  |
| -1 | We [China and America] are in a race and China has to learn to defend its own interests.  | Indeed, Japan has many strengths that attract many Chinese. But don’t forget the national humiliation during Japanese invasions.  | We need to grow stronger and better and make Taiwan people want to come back. | Taking Diaoyu back is a strategic move that requires much thought and effort. We shouldn’t rush.  | This is a critical moment for China. We have to launch political reform, or we will be abandoned by the world.  |
| 0 | America has a lot of problems, so does China.  | I know about the Nanjing massacre, but I still like Japanese movies and sushi. Is it conflicting? No. | Taiwanese people live such a peaceful and happy life. But they hold a feeling of superiority, which makes me annoyed. |  I care much more about my dinner than Diaoyu. It’s too far away from my life.  | Be realistic people. Each system has its own pros and cons. We’d face it and fix it. |
| 1 | The US has been a sole superpower in this world. And this is not without reason. We have to learn a lot from them.  | I am not saying that we should forget the pains of wars, but if we really want peace, we should let go of hatred. | Taiwan is an example of successful democratization for Chinese people and we should regard it as a precious gift.  | Diaoyu is not worth the price of our economic and social development.  | Democracy should be based on a prosperous economy. We will realize democracy when we are strong and rich. But now? We will only be the next India.  |
| 2 | America is my favourite because it has the best two things in the world: democracy and freedom.  | China repeatedly denounces Japan but who cares? Japan is still a country respected and loved by a lot of people around the world. And China is not.  | Nobody will allow an autocratic regime to unify with a democratic regime. This is evil.  | Chinese who want to defend Diaoyu are so pathetic. They can’t even defend themselves in the face of the state.  | Each country is so different that universal values could be an illusion. One-party rule has its unique advantage. Don’t be brainwashed by Western values.  |

*Note:*

All examples given have been drawn from real data, but we have combined and paraphrased typical tweets in each category so as to give a clear illustration of how we assigned sentiment scores, as well as to protect data anonymity.

Appendix Table 5: **Inter-coder Reliability Statistics**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Topics** | **US** | **Regime** | **Japan** | **Taiwan** | **Diaoyu** |
| Agreement Rate | 86.84% | 86.23% | 90.63% | 84.32% | 79.42% |
| Cohen’s Kappa | 0.66 | 0.61 | 0.73 | 0.64 | 0.51 |
| Spearman’s rho (at 0.05 level) | 0.88 | 0.73 | 0.85 | 0.79 | 0.67 |

Appendix Table 6: **Distribution of Censored Tweets in 2012**

|  |  |  |
| --- | --- | --- |
| **Month** | **Population Count** | **Sample Count** |
| January | 120 | 35 |
| February | 2,658 | 35 |
| March | 4,958 | 50 |
| April | 2,981 | 35 |
| May | 3,861 | 39 |
| June | 3,652 | 37 |
| July | 8,853 | 89 |
| August | 2,490 | 35 |
| September | 6,883 | 69 |
| October | 6,319 | 63 |
| November | 13,451 | 135 |
| December | 29,858 | 299 |
| TOTAL | 86,083 | 921 |

Appendix Table 7: **Inter-coder Reliability Statistics for Coding Censored Weibo Tweets**

|  |  |
| --- | --- |
| **Statistics** |  |
| Agreement Rate | 80.67% |
| Cohen’s Kappa | 0.723 |
| Spearman’s rho  | 0.804 |

Appendix Table 8: **Coding Scheme for Censored Weibo in 2012**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Categories** | **Illustrations** | **Examples** |
| 1 | anti-regime & non-nationalist | being critical of the government and/or the political system *without* nationalistic concerns | “Why are there stars on our national flag? Because we are living in darkness. Why is the sun on Taiwan’s national flag? Because they are living in the light.”  |
| 2 | anti-regime & nationalist | being critical of the government and/or the political system for nationalistic reasons | “The state policies are heavily influenced by other countries and so-called democracy and human rights. But this is our domestic affair!” |
| 3 | not anti-regime & nationalist | not being critical of the government, but expressing nationalist sentiments | “Down with Japanese devils!” |
| 4 | not anti-regime & not nationalist | not being critical of the government, and not related to nationalist topics. | [1] “Today a big fire in a Tianjin supermarket caused large causalities.” [2] “Did Bo Xilai attempt to launch a coup?” |
| 5 | other | porn, spams, or tweets without sufficient information  | [1] “Retweet this post to win 1 million award!”[2] “Retweet.” (without the original tweet) |

Appendix Table 9: **Estimated Percentage Range of Each Category in 2012 Censored Weibo Tweets**

|  |  |  |  |
| --- | --- | --- | --- |
| **Categories** | **Sample count**  | **Sample percentage** | **99% confidence interval**  |
| 1 | 340 | 36.92% | (32.81%, 41.02%) |
| 2 | 4 | 0.43% | (0%, 0.99%) |
| 3 | 8 | 0.87% | (0.08%, 1.66%) |
| 4 | 284 | 30.84% | (26.91%, 34.76%) |
| 5 | 285 | 30.94% | (27.01%, 34.87%) |
| **TOTAL** | 921 | 100% |  |

1. Based on the authors’ local knowledge, we included both left-leaning and right-leaning users who are famous political or social commentators. Examples familiar to China specialist include, for example, Ren Zhiqiang 任志强, Sun Liping 孙立平, Wuyue Sanren 五岳散人, Zuoyeben 作业本, Sima Nan 司马南, Hu Xijin 胡锡进, Dai Xu 戴旭, Cai Xiaoxin 蔡小心. [↑](#endnote-ref-1)
2. TF-IDF, an abbreviation of Term Frequency–Inverse Document Frequency, is a statistic widely used in text mining and information retrieval which combines two weightings: the term frequency of a word in a document, and the inverse document frequency of this word. The latter helps to adjust for the fact that some words appear frequently in general but are less important than other rarer words. [↑](#endnote-ref-2)
3. Note that we removed non-individual accounts, e.g. brands, media, and institutes, as we are interested in individuals rather than organizations. [↑](#endnote-ref-3)
4. Note that this number drops to 146 when we zoom in to study nationalism. See discussions in section 2. [↑](#endnote-ref-4)
5. The Word2Vec utilizes the skip-gram architecture and is trained with the full 2013 Weibo corpus of 2.7 billion tweets. In a nutshell, Word2vec projects unique words in the large corpus to a high-dimensional space, such that words sharing common contexts are located in close proximity to one another in the space. [↑](#endnote-ref-5)
6. The crackdown on opinion leaders occurred in late 2013 but was aimed at silencing anti-regime opinion leaders, not nationalists. Since the policy environment remains the same for nationalists, it is still reliable to use the 2012 data to infer censorship of nationalists in 2013. [↑](#endnote-ref-6)