Real-world Islands in a Social Media Sea: Nationalism and Censorship on Weibo during the 2012 Diaoyu/Senkaku Crisis*

Christopher Cairns[†] and Allen Carlson[‡]

Abstract

During August and September 2012, Sino-Japanese conflict over the Diaoyu/Senkaku Islands escalated. Alongside street demonstrations in China, there was an outpouring of public sentiment on China's leading micro-blog, Sina Weibo (微波). Using human and computer-assisted content analysis, we exploit original Weibo data to measure how public sentiment in China fluctuated over the dispute, and ask two questions. First, how cohesive and volatile were online nationalist sentiments? Second, we measure government censorship of Weibo in order to ask which sentiments did authorities allow to be expressed, and when? We first find that many of the micro-bloggers' harshest invective was directed not at Japan but at their own government. Second, while censorship remained high across topics for most of the dispute, it plummeted on 18 August – the same day as bloggers' anger at Beijing peaked. These observations suggest three theoretical explanations: two are instrumental-strategic ("audience costs" and "safety valve") and one is ideational (elite identification with protesters).

Keywords: social media; China; Weibo: nationalism; censorship; Diaoyu Islands; Japan

During the summer and autumn of 2012, an escalation of the Sino-Japanese conflict over the sovereign status of the Diaoyu/Senkaku Islands (*diaoyudao* 钓鱼岛) unfolded in East Asia. The crisis, which grew out of competing Chinese and Japanese claims to the territory, began in mid-August when a small group of protesters set sail from Hong Kong with the intent of landing on the disputed terrain. Upon reaching their destination, they were immediately detained (and later released) by the Japanese. Their detention infuriated many within China.

* We would like to thank King-Wa Fu, Chung-Hong Chan and Michael Chau at the University of Hong Kong for their efforts on the *WeiboScope* project, which produced the dataset we use. We would also like to thank Jing Li, who served on our coding team.

[†] Cornell University. Email: cmc467@cornell.edu (corresponding author).

[‡] Cornell University.

© The China Quarterly, 2016 doi:10.1017/S0305741015001708 First published online 19 January 2016

https://doi.org/10.1017/S0305741015001708 Published online by Cambridge University Press

Subsequently, on 11 September, Tokyo nationalized some of the islands by purchasing them from private Japanese owners. This move may have been designed to de-escalate the tense stand-off over this maritime region; however, if so, it wildly backfired as it was widely viewed within China as an especially provocative measure. As such, it catalysed a series of mass Chinese protests against Japan. Many of these demonstrations were accompanied by harsh anti-Japanese rhetoric and attacks against property and individuals perceived to have ties to Japan.

This article grows out of the conviction that studying how this dispute was portrayed on China's leading social media site, Sina Weibo (*weibo* 微博), offers us a vantage point from which to contribute to the expanding academic debate about both Chinese social media and Beijing's approach to the disputed islands. The paper draws on a new big dataset of over 40,000 influential bloggers on Weibo and measures micro-bloggers' sentiments during the 2012 crisis.[1] We isolated the posts within this collection that were written in August and September 2012, then culled these down to those topically relevant to the Diaoyu/Senkaku Islands and used a combination of hand-coding, keyword analysis and computer-assisted text analysis (CATA) to categorize the sentiments expressed in these posts and identify how such expressions changed over time. This process gives us the most comprehensive study to date of how Weibo users discussed the islands as China and Japan were squaring up in the sovereignty conflict, and allows us to ask the following questions. First, what was the degree of cohesiveness, and level of volatility, of nationalist sentiments such as those that drove forward the mass protests? Second, were such sentiments allowed to be expressed, or were they censored? If the latter, to what degree were they censored, given that these sentiments unfolded simultaneously with real-world street demonstrations and in the context of the Chinese state's extensive online censorship programme? Conversely, if such virulent views were not censored at certain times, why not?

To preview our findings, we first note that there is significant heterogeneity among Weibo users in terms of the sentiments they express. Surprisingly, the referent of many micro-bloggers' harshest nationalist invective was directed not towards Japan but at a Chinese state they characterized as ineffectual and corrupt. This level of anti-government sentiment was both deeper and broader than we had anticipated. However, we also discovered that virulent voices did not always dominate Weibo traffic. In fact, moderate posts, especially those subsequent to crucial real-world events in September, served as a counterpoint to the more rabid views that until then had held sway. During the periods of greatest Weibo use, such level-headed commentaries actually eclipsed the more virulent sentiments and revealed that, at least in this particular case, the forum was not as monopolized by angry youth as conventional wisdom would lead one to expect.

[1] The data are available at: http://147.8.142.179/datazip/.

https://doi.org/10.1017/S0305741015001708 Published online by Cambridge University Press

This finding then leads us to question why such comments were not uniformly censored for the entire duration of the dispute, given the Chinese government's strategy of suppressing comments that "represent, reinforce, or spur social mobilization."[2] Specifically, we develop a method for estimating the proportion of all Weibo posts in our sample that were censored on each day of the dispute (roughly 49 days) between August and September 2012. We find that while censorship was consistently high in September, it dropped precipitously in mid-August immediately after the protesters landed on one of the islands and during an initial wave of street demonstrations.

Our data strongly suggest three theoretical explanations that address the above questions: two are based on "instrumental-strategic" logics, and one on an ideational or "belief-driven" background condition. First, our finding that censorship plummeted in mid-August is strongly consistent with the logic of "strategic censorship,"[3] a phenomenon posited by scholars of anti-foreign protests[4] and Chinese nationalism more specifically,[5] but which until now has lacked strong empirical testing. This logic has both domestic and international components. Domestically, once China's leaders sensed that the Chinese public was enraged, they realized the need for a "safety valve"[6] through which such anger could be "vented" (*faxie* 发泄) in order to prevent it from leading to greater social instability or turning even more sharply against the state. Internationally, Beijing had few strategic options for showing its determination to defend China's territorial claim, and it appears that allowing virulent online opinion to be expressed, along with its harsh official rhetoric to generate "audience costs,"[7] only served to tie its hands further in regard to its dispute with Japan.

While these two explanations provide the most leverage to explain exactly when and under what circumstances the drop in censorship occurred, alone they are arguably insufficient to account for why China's leaders would be willing to tolerate even a moderate amount of risk in the first place in order to carry out such an instrumental strategy, given that they routinely suppress all manner of collective actions and related speech, even in cases where such mobilization does not obviously threaten social stability or state legitimacy. To explain why the anomaly of large-scale nationalist protest in China and the accompanying internet commentary, which is not permitted to the same degree on any other issue, is even thinkable among Chinese elites requires understanding the extent to which they have become predisposed to identify with the nationalist rhetoric they have espoused as a legitimizing narrative since the 1989 Tiananmen crisis. In other words, China's leaders were primed to look favourably on protesters' anger at Japan over perceived provocations during the dispute. This then made

[2] King, Pan and Roberts 2013, 1.
[3] Lorentzen 2014.
[4] Weiss 2013.
[5] Reilly 2011.
[6] Hassid 2012.
[7] Weeks 2008.

https://doi.org/10.1017/S0305741015001708 Published online by Cambridge University Press

thinkable the possibility of temporarily allowing such animosity to be expressed online, even when it was directed back at Beijing itself, in an exception to the usual repression of information to which such mass mobilizations in China are subject.

Therefore, although our data speak most strongly to the instrumental nature of the leaders' decision to ease censorship on 18 August and cannot show that such belief-driven considerations were a proximate cause of the pattern we observe, we also caution against overlooking the elites' own national identity construct as a necessary background condition that motivated them to find a way to express their own anger against Japan – while working within the "rationalist" calculus of balancing risk to social or regime stability with the tangible benefits of sending strategic signals to domestic and foreign audiences. Chinese leaders may well have bought into the nationalist narratives they themselves had propagated, and yet, within this overall belief system, acted instrumentally in their handling of the Weibo commentaries.

Finally, although our data do not allow us to identify conclusively a single "correct" explanation, we are still able to rule out two alternatives. First, we provide evidence that such expressions of anger on Weibo were authentically grass-roots. They emerged from the bottom-up in direct response to real-world events and were not part of a public relations campaign orchestrated from Beijing. Second, our empirical findings and theoretical explanations are both largely consistent with the most significant recent findings in the literature regarding online censorship and collective action that contend that entire topics are censored at high rates whenever they closely accompany real-world collective action events.[8] However, our data's overall pattern, and our argument regarding what has shaped it, especially the observed drop in censorship in mid-August, does raise questions about the emerging conventional wisdom in the field about such trends.

Between a Rock and a Hard Place: The Dual Rise of Nationalism and Social Media

One of the core questions for any study of the 2012 Diaoyu/Senkaku Islands crisis concerns the role that Chinese nationalism played in this volatile stand-off. To what extent did nationalist sentiments during this period run amok, and did they reach such a fevered pitch that they came close to pushing Beijing into armed conflict with Japan? Such a query, while provocative, constitutes nothing more than an extension of a long standing debate among students of contemporary Chinese nationalism about the degree to which such a collective identity is on the rise in China, and who has agency – the state or the people?

These inquiries were first forwarded by leading Sinologists such as Allen Whiting and Michel Oksenberg.[9] They were then taken up by a handful of

[8] King, Pan and Roberts 2013, 2014.
[9] See Whiting 1983; Oksenberg 1986.

https://doi.org/10.1017/S0305741015001708 Published online by Cambridge University Press

pioneering scholars, including Suisheng Zhao and Peter Gries, amongst others.[10] All of this work tended to agree that new Chinese nationalist narratives drew upon collective memories of China's historical greatness during the dynastic period, the country's suffering under foreign aggressors during the "century of humiliation," and its subsequent rise under the leadership of the Chinese Communist Party (CCP). Beyond such points of agreement, however, earlier observers were divided over the balance within such stories between the belief-driven (in terms of their resonance at a popular level and the degree to which they were internalized within China) and the instrumental (as intended to inculcate loyalty and secure state legitimacy).

Nearly a decade of subsequent scholarship has repeatedly portrayed Chinese nationalism as a double-edged sword, which China's leaders both manipulate to their advantage and are periodically beholden to (as it is somewhat beyond their control).[11] However, we find this work wanting on two fronts. First, work to date has focused on events that occurred prior to the rise of social media or, when analysing more recent developments, it has not placed social media at the centre of analysis. At most, one sees an inclusion of various facets of Chinese cyberspace in analysis of Chinese nationalism, but this has largely been limited to looking at websites and blogs and not social media. Yet, it is clear that social media have become one of the most popular forms of political internet usage in China. More pointedly, the 2012 Diaoyu/Senkaku confrontation was China's first major post-Weibo international crisis, and the crisis was one of Weibo's top trends of 2012.[12] This, then, begs analysis that gives this new medium pride of place.

Even more importantly, recent work has done little to shed new light on the overarching question of the extent to which Chinese nationalism has become a Pandora's box for the Chinese state, one that it opened long ago, and can no longer close. There are two primary reasons for such an oversight. First, when students of Chinese nationalism have talked about nationalism as a threat, or a risk, to Beijing, they have done so via the use of conceptually vague language. For example, Jessica Weiss's *International Organization* article, which treats managing nationalist protests as a precarious balancing act for China's leaders, notes that, "In managing nationalist protests, autocrats weigh the risk to the status quo against the cost of using force or coercion to prevent citizens from gathering in the street."[13] In the same vein, James Reilly references danger to the state, noting, "For Chinese leaders, swelling nationalist protests pose a dangerous dilemma."[14] In addition, Suisheng Zhao's recent re-appraisal of Chinese nationalism contends that,

[10] Zhao 2004; Gries 2004.

[11] Hughes 2006; He 2007; Reilly 2011; Rozman 2013; Weiss 2013.

[12] Sina Weibo was launched in 2009 but only took off in popularity during 2011. Its popularity surged after the 2010 incident involving a collision between a Chinese fishing trawler and two Japanese coastguard vessels.

[13] Weiss 2013, 7.

[14] Reilly 2014, 200.

https://doi.org/10.1017/S0305741015001708 Published online by Cambridge University Press

"Enjoying an inflated sense of empowerment supported by its new quotient of wealth and military capacities, and terrified of the uncertain future due to increasing social, economic and political tensions at home, the Chinese state has become more willing to play to the popular nationalist gallery in pursuing the core interests."[15]

We suspect that such claims may have validity; however, we are also struck not only by their vagueness but also the extent to which they lack any clear approach for measuring the extent of such apparently rising challenges. To be blunt, if the theoretical danger of nationalism to the Chinese state is left this vacuous, then it is impossible to know how much of this sentiment is percolating within China. Thus, the second limitation of the discussion of Chinese nationalism has been its lack of empirical referents.[16]

In contrast to the literature on Chinese nationalism, the study of censorship in China and elsewhere is only a few years old. Prior to the recent wave of methodologically sophisticated studies,[17] work by authors such as Rebecca MacKinnon[18] and Evgeny Morozov[19] explicated the mechanics and theorized about the intent of China's and other authoritarian governments' censorship programmes. These authors tended to conclude that autocrats were interested in censoring the internet either simply because citizens used online space to criticize officials or policies, or because they used it to organize collective actions. What these authors did not consider is the possibility that, in select instances, the state might choose *not* to censor the internet in order to derive benefit from what citizens were saying. In other words, while recent studies have made great strides in establishing a broad motive for Beijing to suppress online promotion of collective action, we are not satisfied that they have explained all, or even most, variation in precisely when and where censorship occurs.

This article, while not free of its own shortcomings, addresses the defects in both literatures. We set up a "risk" ladder for classifying different strains of nationalist sentiments according to the degree we think Beijing would view them as threatening, and from these are able to make descriptive inferences about the first research question above. To address the second question, we estimate the day-by-day censorship rate during the dispute (the percentage of all posts deleted by censors on a given day).

Weibo as Big Data: Measuring Censorship and Nationalist Sentiment in Real Time

We addressed the questions raised above by utilizing a trailblazing dataset collected by researchers at the University of Hong Kong.[20] To our knowledge, it

[15] Zhao 2013, 536–37.

[16] Notable exceptions include Shen 2007; Callahan 2009; Johnston and Stockman 2007; Carlson 2009; Gries et al. 2011; Cheng 2011; Hoffman and Lerner 2013.

[17] King, Pan and Roberts 2013; Zhu et al. 2013; Bamman, O'Connor and Smith 2012.

[18] MacKinnon 2012.

[19] Morozov 2011.

[20] See Fu, Chan and Chau 2013 for full citation. Data accessible at: http://147.8.142.179/datazip/.

https://doi.org/10.1017/S0305741015001708 Published online by Cambridge University Press

is the most comprehensive dataset of Weibo posts currently available. The methodology used to collect it is described in detail by Fu, Chan and Chau.[21] Each row in the dataset consisted of one social media post (including reposted or "retweeted" content), plus associated meta-data. The data files did not include multimedia such as images or videos. We relied only on the post text and counted embedded reposts as part of the text.[22]

To prepare the data for analysis, we first extracted posts placed between 13 August and 30 September, as we found these contained the vast majority of Diaoyu-related commentary. Next, we filtered the posts according to whether they contained a keyword, "Diaoyu Islands" (*diaoyudao* 钓鱼岛), that was highly predictive of discussion of the dispute. This left about 145,000 posts over 49 days. One immediate concern with filtering using this (or any) keyword is whether estimates of sentiment categories derived from such a sample are generalizable to the broader population of all topic-relevant posts. Accordingly, we present select results below for both posts containing "Diaoyu Islands" and a sample that did not but which we found, through careful reading, also to be topically relevant. To preview our findings, adding this additional sample strengthens our overall results.

How dangerous a notion? Making sense of how "risky" Weibo posts appear to the Chinese state

Beyond simply tracking how much Weibo commentators said during this period, we were especially interested in determining the target of any such commentary. To structure our inquiry in this regard, we developed a set of sentiment categories to apply to the data. The process was iterative. First, we carefully read through a random sample of posts drawn from across the time period in order to sort the posts inductively into categories based on their distinctive sentiment strains, alongside a few "residual" categories such as reposts of news content. We then organized the non-residual categories into an ordinal "collective action risk ladder," ranging from one to five. With this scale established, the two authors, along with a third coding assistant, worked to assign a separately drawn random sample of 479 posts into one of the eight categories. After working independently, the three coders met to reconcile scores, and we report inter-coder reliability statistics in Appendix A.

The key to this enterprise was the "risk ladder" that hand-coding allowed us to operationalize. At the lowest rung (level one), we designated as "moderate" those posts which cautioned against the potential negative consequences of direct action against Japan, or which objected to the violent turn that protests took over the course of our 49-day sample. The latter of these views is exemplified by the statement: "Protecting the Diaoyu Islands does not involve harming our

[21] Fu, Chan and Chau 2013.
[22] See Appendix B for more detail.

https://doi.org/10.1017/S0305741015001708 Published online by Cambridge University Press

fellow citizens!" (*baowei diaoyudao bushi shanghai ziji de tongbao* 保卫钓鱼岛不是伤害自己的同胞!). As such posts mostly opposed the virulent and even violent excesses of more extreme fellow Chinese, we felt that Beijing would view these voices as posing little risk.

The second category (level two) included posters who parroted Beijing's own official statements about the Diaoyu Islands, and in particular their status as part of China, and the country's determination not to cede such territory. While such a category could encompass a wide variety of phrases, we anticipated its most common refrain to be the patriotic assertion: "The Diaoyu Islands are China's" (*diaoyudao shi Zhongguode* 钓鱼岛是中国的). Such posters were at least partially mobilized. Therefore, they represented an *a priori* risk to Beijing compared with the moderates. However, this risk was minimal because the sentiments these individuals endorsed were so orthodox vis-à-vis the state.

The third rung (level three) encompassed posts containing strongly anti-Japanese sentiments and those that actively blamed the Chinese people (not state) for their weakness in not standing up to Japan. The "anti-Japanese" subset of this category generally spouted derogatory anti-Japanese rhetoric, such as "little Japs" (*xiao Riben* 小日本), "Japanese pirates" (*wokou* 倭寇), or "Japanese devils" (*Riben guizi* 日本鬼子). In contrast, the second subset contained posts lamenting China's inferiority or fecklessness vis-à-vis Japan *without* in any way implicating Beijing. While these two subsets are substantively very different, we estimated that they posed approximately the same degree of risk to the state. The reason that this group ranks as a higher risk than the level two group is owing to the fact that posters' sentiments about the dispute went well beyond the scope of the state's official stance on the islands.

The next ordinal category (level four) consisted of posts containing specific calls to action. These included, for example, endorsing "boycotting Japanese goods" (*dizhi Rihuo* 抵制日货), supporting "protests" (*kangyi* 抗议), or even voicing demands to "deploy military forces" (*pai bing* 派兵). Not surprisingly, we found this category to be quite heterogeneous. Such sentiments called for more action than Beijing had taken, and were more pointed than the invective that predominated at level three as they contained specific suggestions and/or threats regarding how best to settle China's score with Japan. As such, these posts increased the risk of collective action for the government, as we believe individuals willing to vocalize the desire to "do something" would be more likely, all other things being equal, either to join street demonstrations or endorse them online.

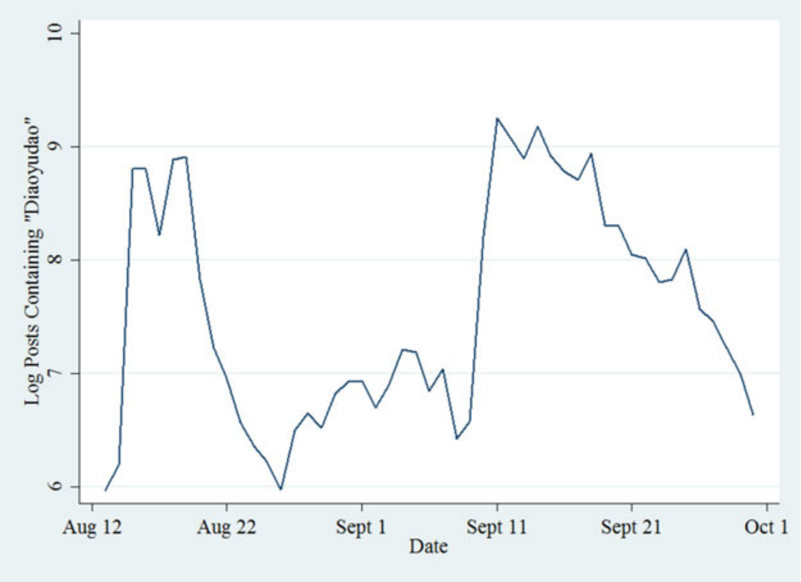
Finally, we anticipated that some posts (level five) would pose a high risk to China's leaders in that they would directly criticize the state (either for its handling of the dispute, or along broader lines). Posts that darkly referred to the abuses of power of China's notorious City Urban Administrative and Law Enforcement Bureau (*chengguan* 城管) – cemented in the popular consciousness as low-level police "thugs" – or that highlighted the impotence of China's "dear

https://doi.org/10.1017/S0305741015001708 Published online by Cambridge University Press

military" (*guibing* 贵兵) and/or the Communist Party, were all conceptualized as part of this category.

After developing the categories and completing hand-coding, we paused to view a simple summary statistic: post counts of *diaoyudao* for each of the 49 days. We calculated these directly from the population of 145,000 posts to determine if there were sub-sets of days that merited particular attention. Figure 1 shows these results.

Figure 1: **Daily Posts Containing "Diaoyu Islands"**



Looking at the graph, we noticed four dates (16 and 18 August, and 11 and 15 September) that stood out as having high post volume and that corresponded to important real-world events. The first date, 16 August, was the day after activists landed on one of the Diaoyu Islands, and captures much of netizens' initial reaction to events. On 18 August, there was a peak of public criticism of the Chinese government's perceived inactivity in securing the activists' release from Japanese authorities. This was also the day before a group of Japanese activists landed on the islands. On 11 September, the Japanese government officially nationalized the islands, and 15 September was the beginning of the weekend that saw the largest street demonstrations of any weekend during the time period. While we could have also picked other key dates of interest, we believe that these four – two from each wave – adequately capture the main peaks in netizens' reaction to events. We therefore subjected these dates to closer scrutiny, and drew four additional random samples of 150 posts, one for each date.

https://doi.org/10.1017/S0305741015001708 Published online by Cambridge University Press

So, what was being said? Keywords as indicators of sentiment

Hand-coding post samples for the entire time period and for select days allowed us to estimate the proportions of our sentiment categories for the whole period as well as for those days. However, we wished to go beyond this and to infer the proportions for other, non-sampled days, thus allowing us to generate time series consisting of daily category proportions for the entire period. We used two measurement techniques towards this end. First, we used keywords to proxy for what we found, through extensive reading, to be particularly characteristic refrains in each category, and second, we used a computer-assisted text analysis (CATA) method called *ReadMe*.[23]

We leave discussion of *ReadMe* to the next section, and here introduce our keyword measures. Beginning with the moderate category, we discovered that two terms, "rational patriotism" (*lixing aiguo* 理性爱国) and "smash" (*za* 砸 – with reference to admonishments not to carry out such activities against Japanese goods owned by Chinese citizens), were the most prevalent terms. Not surprisingly, the simple declaration that "the Diaoyu Islands are China's" was the dominant refrain within our patriotic/parroting category. Owing to a lack of meaningful variation, we ended up dropping level three (anti-Japanese and self-blame sentiments) from the analysis, as it did not seem to relate closely to either current events or the other series. Next, the call to action category was dominated by various statements in support of "boycotting Japanese goods," then underway within China during the late summer and early autumn of 2012. Finally, the highest risk category, while containing a wide array of criticisms, was most consistently voiced via protesters' satirical use of the phrase "heavenly dynasty" (*tianchao* 天朝), a reference to the ineptitude and decline of the Qing dynasty, to refer to the contemporary Chinese state. While on the surface the phrase is one of deep respect and admiration, within the context of our posts it was used to underscore the shortcomings of China's leaders and the domestic political status quo. We derive this reading of the term from the manner in which it was consistently used in unison with sharp criticism of government policies, and also from our review of the secondary literature on Chinese social media, which has recently highlighted the term as a particularly biting form of anti-government commentary.[24]

The Censorship Rate, Sentiment Categories and Weibo's Fluctuating Risk and Benefit during the Crisis

When was the censorship rate "low" and what did this mean? Estimating the rate of Weibo post deletions

Our findings depend on two critical observations: first, that the censorship rate was low during the protest wave in August but high in September, and second, that

[23] Hopkins and King 2010.
[24] Link and Xiao 2013.

https://doi.org/10.1017/S0305741015001708 Published online by Cambridge University Press

which sentiment category in each wave was largest be counterintuitive to observed low or high censorship rates for that wave – in other words, that the prevalence of certain sentiments in each wave be unanticipated by existing theory, given that censorship is high or low. First, we estimate the censorship rate, i.e. the percentage of Weibo posts deleted out of all topic-relevant posts on a given day. To narrow the comparison, we choose to focus on 18 August and 15 September as the peaks of the two waves. Table 1 shows estimated censorship rates for these two dates, both among posts containing *diaoyudao* and posts without this keyword.[25]

Table 1: **Censorship Rates for Select Dates with and without *Diaoyudao* Keyword**

Date	**95% confidence interval (mean): posts without *diaoyudao***	**95% confidence interval (mean): posts with *diaoyudao***
18 August	0–38.5% (0%)	22.3–27.1% (24.7%)
15 September	51.0–82.5% (69.9%)	58.9–62.1% (60.5%)

While confidence intervals for sample estimates are wider owing to our smaller sample of non-*diaoyudao* keyword posts, it is still evident that censorship is much lower on 18 August than 15 September. Taking this a step further, we were able to estimate the censorship rate (for posts containing *diaoyudao*) across the entire time period, as shown in Figure 2.[26] This graph shows that the censorship rate remained fairly constant throughout the dispute except on 18 August, when it plummeted to about 22–27 per cent. This contrast suggests that authorities' ordinary response for much of the dispute was to censor at a high rate – the fact that the rate dropped so sharply in mid-August, therefore, suggests that they made a deliberate decision to censor less at that time.

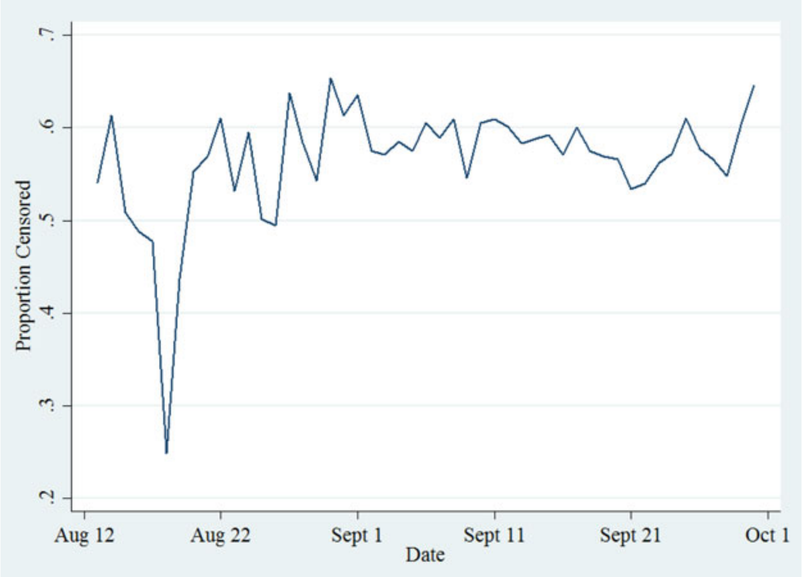
Turning to the estimates for our two key dates, we note that the results for posts lacking *diaoyudao* accentuate those derived from sampling on that keyword. For 15 September, the mean for non-keyword posts is higher than for posts with the keyword. Similarly, while the interval for non-keyword posts on 18 August is large, its lower bound (0 per cent), which is also its mean, lies well below that for keyword posts, raising the likely possibility that censorship on that date was even lower than the mean of about 25 per cent we observed in the keyword sample.

[25] For posts without the *diaoyudao* keyword, we analysed a small sample. Our estimates are thus subject to some sampling variance. For 18 August, we drew a sample of 49 posts, of which none were censored. As no "successes" occurred in this binomial variable, we applied Hanley and Lippman-Hand's (1983) "Rule of Three" to estimate the upper 95% confidence bound. For 18 September, we sampled 51 posts, of which 10 were censored, and calculated the confidence interval from a binomial distribution. In both cases, we adjusted the observed censorship rates in the *WeiboScope* data to account for the data's downward bias. The numbers here rely on Zhu et al.'s (2013) finding that 90% of post deletions occur within 24 hours of an initial event. We assume our data have the same speed of censorship as theirs. See Appendix B for more details.

[26] Figure 2 depicts posts containing *diaoyudao* and relies on the same assumptions as the previous footnote.

https://doi.org/10.1017/S0305741015001708 Published online by Cambridge University Press

Figure 2: **Percentage of Censored Posts over Time**



Moving on to sentiment: not all Weibo commenters were angry, and when they were, their target was not always Japan

Confident that we had estimated the censorship rate with sufficient accuracy and that its dip in August cut against the conventional wisdom, we next turned to estimating sample proportions for all categories, both averaged across the entire time series and for specific days, as shown in Table 2. The 49-day results were estimated directly from our hand-coding of 479 observations, and the day-specific results from the 150 observations drawn for each of the four targeted days. In Table 3, we also again present results with and without the *diaoyudao* keyword.

Table 2: **Categories (*Diaoyudao* Keyword Sample) by per cent of Total for Select Dates**

Category (% posts)	13 Aug–30 Sep	16 Aug	18 Aug	11 Sep	15 Sep
1. Moderate	4	6	2	3	19
2. Patriotic	13	13	6	21	9
3. Anti-Jap/self-blame	10	5	2	7	5
4. Call to action	9	11	21	11	6
5. Anti-government	13	29	51	7	10
6. News	29	20	8	31	23
7. Uncodeable/other	16	11	9	13	20
8. Humour/satire	5	5	1	8	7

https://doi.org/10.1017/S0305741015001708 Published online by Cambridge University Press

Table 3: **Categories for 18 August–15 September with and without *Diaoyudao* Keyword**

Category (% posts)	**18 Aug w/keyword**	**18 Aug w/o keyword**	**15 Sep w/keyword**	**15 Sep w/o keyword**
1. Moderate	2	2	19	45
2. Patriotic	6	4	9	0
3. Anti-Jap/self-blame	2	10	5	12
4. Call to action	21	10	6	14
5. Anti-government	51	71	10	6
6. News	8	0	23	8
7. Uncodeable/other	9	2	20	14
8. Humour/satire	1	0	7	2

While proportions for posts containing, and lacking, the *diaoyudao* keyword differ somewhat, the major trend of each wave – the rise and fall of the "anti-government" and "moderate" categories – is similar for both: the proportion of anti-government posts surges in August, peaking on 18 August at above 50 per cent (and as high as 71 per cent for non-keyword posts). Similarly, the moderates category shows a sharp increase in September, with the trend more prominent in non-keyword posts.

Turning to the other categories, we first note that the patriotic category differs between the two samples on 15 September. This is owing to the fact that since the non-keyword sample excluded posts containing *diaoyudao*, it missed what was by far the most common instance of this category: the phrase "the Diaoyu Islands are China's." This suggests that the patriotic category consisted mainly of people content merely to repeat the above meme and to show their support for China's claim to the islands. What these individuals were *not* doing was either calling for real-world actions or criticizing their own government. Similarly, the news category is lower on both dates in the non-keyword sample. This reflects the fact that many uses of the *diaoyudao* term were embedded in news content that Weibo users then reposted.

Our initial sampling on this keyword thus biased the proportions towards patriotic parroting (per the above phrase) and towards news content, and likely correspondingly reduced the proportions in other categories. Still, our main finding holds up well: anti-government commentary dominated in August and a moderate backlash occurred in September. These two categories, for their respective time periods, are the highest of any of the five ordinal categories in both samples.

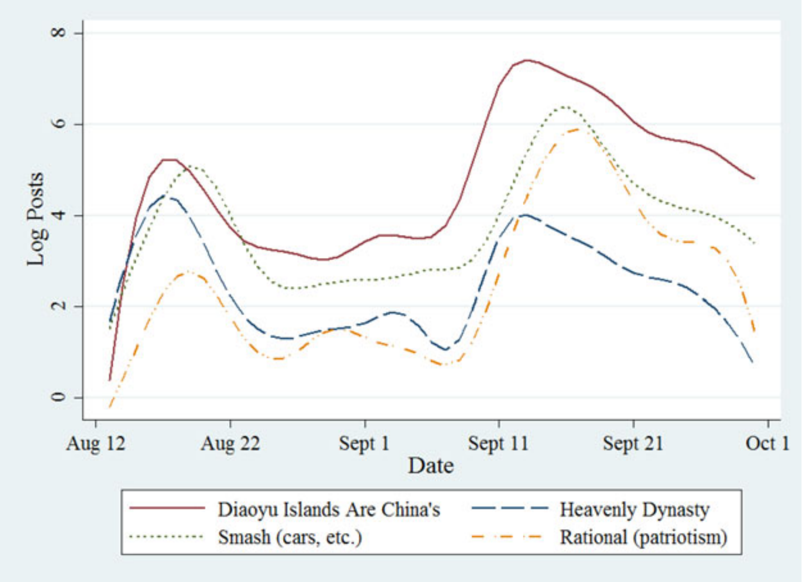
Using more sophisticated measurement techniques to examine Weibo sentiment more closely: still less anger than one might expect, yet more anti-government sentiment than Beijing would have liked

While the above category proportion estimates give a general idea of the timing of key sentiment waves, we wish to explore in greater depth the ebb and flow of

https://doi.org/10.1017/S0305741015001708 Published online by Cambridge University Press

key refrains we found to be characteristic of these categories. Figure 3 shows these for four key categories.[27] The graph's most distinctive aspect is the sequencing of the four time series. The patriotic refrain, "the Diaoyu Islands are China's," as Sina Weibo's tenth most common "hot topic" (*remen huati* 热门话题) of 2012, proxies for micro-bloggers' general attention to the dispute, and particularly their response to real-world events. This phrase surges immediately following the activist landing on 15 August, and actually anticipates by a few days Japan's nationalization of the islands on 11 September.[28] The other keyword series then unfold in the context of this macro trend. "Heavenly dynasty" peaks on 18 August, while two phrases representing a moderate backlash – decrying the "smash[ing]" of Japanese cars and calling for "rational" patriotic expression – surge in mid-September.

Figure 3: **Log Post Counts of Four Keyword Series**



In interpreting these keyword series, we emphasize that they capture only a portion of the variation in each category – a wide range of phrases were used throughout the dispute to express anti-government and moderate sentiments. To address this issue, we used the *ReadMe* program mentioned above, which used the information contained in our coding team's manual classification of posts to estimate category proportions from among all posts containing the

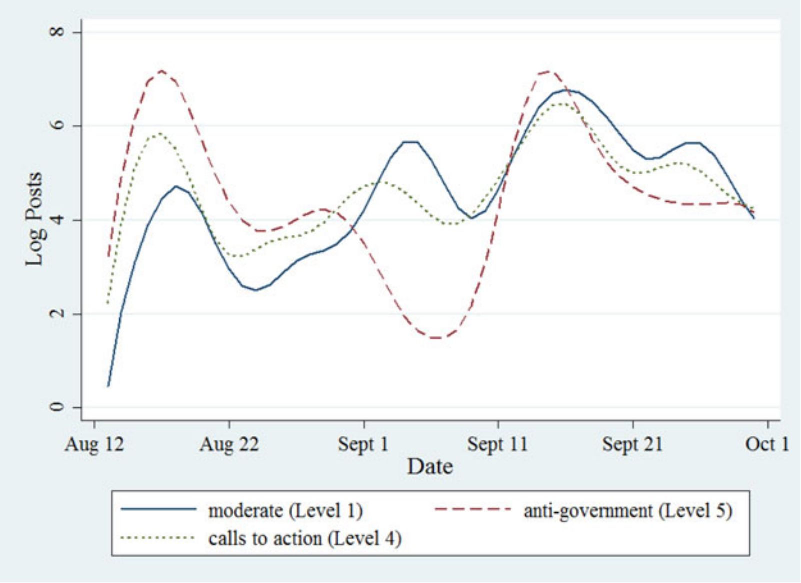
[27] To ensure topic relevance, we constructed these series only from posts containing *diaoyudao*. The graphs are smoothed with second-order polynomials.

[28] Data gathered from Weibo data center at: http://tech.sina.com.cn/i/2012-12-19/13447902817.shtml.

https://doi.org/10.1017/S0305741015001708 Published online by Cambridge University Press

diaoyudao keyword.[29] Figure 4 displays the time series that resulted from this algorithm.

Figure 4: ***ReadMe* Estimates of Category Post Quantities**



While *ReadMe* produced results that diverged from our directly estimated proportions for key dates, the fact that, despite such a discrepancy, the results nonetheless evince the same general trends as our keyword series in Figure 3 gives us confidence that by using multiple measurement techniques we have adequately characterized the timing and rough magnitude of real-world online sentiments during the dispute. For example, we note that anti-government sentiment is highest around 18 August, while moderate views trended upward throughout the dispute and prevailed in mid-September.[30]

[29] We did not run *ReadMe* on our two day-specific (18 August and 15 September) samples of dispute-relevant posts without the *diaoyudao* keyword. However, as we have already established using other measurement techniques that key trends critical for our argument hold in both keyword and non-keyword samples, this poses no issue. For space reasons, we omit discussion here of technical aspects of CATA using *ReadMe*. This information is available upon request.

[30] Also of note is the fact that in both Figures 3 and 4, moderate sentiments surge in late August in addition to their larger wave in September. While to support our theoretical argument we choose to emphasize the moderate surge in September (in contrast to anti-government commentators' dominance in mid-August), an uptick in late August moderate posts does not contradict our interpretation; the August moderate surge occurred *after* 18 August and therefore during a period in which the government re-implemented high censorship, an observation entirely consistent with our argument. Similarly, the presence of a brief surge in anti-government posts around 10 September, just before the Japanese government nationalized the islands, does not challenge our interpretation. The government censored *all* forms of posts at higher rates in September than during the 18 August dip; we specifically focus on the authorities' censoring of moderate posts during this period to sharpen the contrast with August.

https://doi.org/10.1017/S0305741015001708 Published online by Cambridge University Press

Conclusion: Given the Volatility of Expressed Sentiments, Why Was Weibo Not Heavily Censored during the Crisis?

Our most surprising empirical results from this project were threefold. First, the censorship rate of online discussion was lower at a critical juncture in August than extant literature would predict. Second, much of the netizens' anger was directed at Beijing, not Tokyo. Finally and paradoxically, we found that more moderate voices appeared to counter these virulent sentiments directly after a broader swath of the public tuned in to online discussion of the dispute in September.

With these results in hand, we turn to jointly interpreting the facts of low/high censorship at key moments in August/September, the prevalence of certain sentiment strains at these points in time, and the real-world events – both on the Chinese streets and between the Chinese and Japanese governments – that were then unfolding. Beginning in August, the Chinese activists who landed on the islands were detained on 15 August and released two days later. Initial reports of street protests in China broke on 15 August.[31] On 17 August, Japan's Kyodo news service reported that messages had appeared on Weibo calling for anti-Japan demonstrations, and noted that the posts were not immediately deleted.[32] On 19 August, a group of Japanese activists landed on the islands. Perhaps anticipating, or at any rate quickly reacting to this development, protests unfolded that same day in several major Chinese cities and were covered by Xinhua.[33] By 21 August, however, mainstream media were attempting to calm protesters and avert further violent escalation.[34]

The picture that emerges from this chain of events is that Beijing had to have been well aware, from the moment the Chinese activists landed on the islands, if not before, that nationalist-minded Chinese would seize quickly on the activists' detention and immediately begin criticizing their own leaders for showing insufficient resolve. In spite of this, leaders allowed Weibo traffic in precisely this vein to manifest alongside growing street demonstrations. Although the anti-government comments we analysed from 18 August mostly did not reference *specific* collective actions, and thus our findings do not directly contradict those of King, Pan and Roberts,[35] our observation is still puzzling as intuitively we would expect the Chinese state to suppress *all* related discussion on such a sensitive topic, especially when real-world collective actions were already underway.

[31] "Chinese group protests in front of Japan embassy to demand disputed isles," Kyodo News Service (reprinted by BBC Monitoring Asia Pacific), 15 August 2012.

[32] "Internet messages in China call for anti-Japan protests on Sunday," Kyodo News Service (reprinted by BBC Monitoring Asia Pacific), 17 August 2012. That same day, Kyodo reported that protesters in Shenzhen smashed Japanese-branded cars and broke into Japanese restaurants.

[33] "Protests in China against Japanese activists' visit to disputed islands," Xinhua (reprinted by BBC Monitoring Asia Pacific), 19 August 2012.

[34] "Bid to calm public after anti-Japanese protests in Shenzhen over Diaoyus: media praise patriotism over disputed islands, but some call Shenzhen behaviour 'shameful'," *South China Morning Post*, 21 August 2012.

[35] King, Pan and Roberts 2013.

https://doi.org/10.1017/S0305741015001708 Published online by Cambridge University Press

This finding stands in even sharper relief when we consider that, in mid-September, the government censored comments that *condemned* excessive actions. We believe that the theory of collective action potential, supported by King, Pan and Roberts,[36] adequately explains this, as these moderate posts were referencing actual destructive events such as car-smashing, albeit to criticize such acts. If this theory held in September, however, then why not on 18 August? Still more puzzling is the fact that while it suppressed criticism of car-smashing on Weibo, the government continued to tolerate massive street demonstrations across mainland China through 18 September. Moreover, Chinese Foreign Ministry spokesman Hong Lei 洪磊 explicitly mentioned the anti-Japan demonstrations on as late as 19 September as justified in the defence of Chinese sovereignty, while Xi Jinping 习近平 called the islands' purchase a "farce."[37] The *People's Daily* and other major Chinese media outlets also published commentaries which ranged from voicing "sympathy" for the demonstrations to outright support.[38]

In light of the Chinese government's tolerance of or even support for the demonstrations in both August and September, why would it relax control over social media in the first wave but not in the second? What purpose might be served in furthering Beijing's strategic priorities by selectively opening the floodgates of online criticism in China against the government while managing street demonstrations and print media differently? As stated in the introduction, we hypothesize that a combination of three forces, two instrumental and one ideational, underlay this turn of events.

First, the speed with which street demonstrations unfolded after the activists' detention and the simultaneous presence of anti-government sentiment on Weibo are together consistent with the international circumstances. Chinese leaders needed to signal their resolve both to their own people and to foreign audiences to prevent Japan from making further advances to consolidate control over the disputed territory without resorting to military action. These circumstances prompted Beijing to attempt to turn micro-bloggers' anger at their own government to its advantage. Our finding thus supports Weiss's argument that the Chinese state selectively allows nationalist protests as a form of risky behaviour that pressures Beijing to stand firm in international disputes, lest the protesters escalate their actions and turn against their leaders.[39] While Chinese leaders in 2012 followed this logic regarding real-world demonstrations, social media platforms such as Weibo are also recognized for their ability to amplify collective action, as shown by the use of Facebook and Twitter in the Arab Spring.[40] Allowing Weibo discussion to take place alongside street demonstrations during

[36] Ibid.

[37] "Xi slams Diaoyu 'purchase'," *China Daily* (European Edition), 20 September 2012.

[38] "Beijing mixes messages over anti-Japan protests," *The New York Times*, 17 September 2012.

[39] Weiss 2013.

[40] Hussain and Howard 2013.

https://doi.org/10.1017/S0305741015001708 Published online by Cambridge University Press

the dispute served as Beijing's intended signal to elevate Japanese (and US) perceptions of the domestic instability risk facing China.

Second, our observations raise the possibility that Beijing felt it had to let nationalist-minded netizens "blow off steam" (*faxie* 发泄). The 2012 Diaoyu dispute was a high-profile event which many netizens quickly learned about. Research suggests that *visible* censorship – censorship that online citizens know is occurring – tends to backfire by inducing greater information-seeking behaviour.[41] In other words, had Beijing resorted to crude information suppression immediately after the activists' detention in an attempt to quell the narrative that it was to blame for being spineless, this action might paradoxically have propelled angry netizens to seek out and push this narrative.

Third, just as the surge in anger at Japan and netizens' simultaneous criticism of Beijing pushed leaders to allow these sentiments to be vented domestically, it is also likely that leaders personally shared the anger expressed by citizens online and in the streets. It would be surprising to think that China's ruling elites were immune to the anti-Japanese nationalist rhetoric which they had regularly produced over the years prior to the 2012 crisis. As such, they were predisposed towards looking favourably upon an online platform that provided them with a mechanism through which anger directed towards Japan could be expressed without the greater destructiveness and risks to social stability that accompany street actions. This is not to imply that the Chinese leaders blindly congregated around such a refrain; rather they lived within a social milieu (of their own making) in which negative perceptions of Japan had become naturalized. Thus, they were inclined to turn a blind eye towards Weibo posters at such a critical juncture in Sino-Japanese relations at a time when Tokyo was widely considered within China to have trampled upon the country's sovereign rights and disrespected its rising international status. This is not to say that leaders in Beijing were governed by their emotions, but rather that their instrumental calculations emerged from a cauldron of xenophobia.

Because the dispute was still in its early stages in August, Beijing may have been more willing to tolerate some genuine risk in order to signal its resolve. By September, however, demonstrations had escalated and become more destructive. While Beijing may not have been able (or willing) to rein in street demonstrations immediately, it was much more capable of tightening online censorship in order to stem the spread of information about such destructive acts. Furthermore, as time passed and more moderate voices entered the online sphere to oppose the virulent excesses of the protesters, the government no longer faced the same pressure to permit an outlet through which extreme voices could vent their anger – they had already done so and would eventually need to be reined in on the streets as well as online. While street protests raged on for another few days and Chinese leaders continued to make tough statements, censoring Weibo represented a first step towards bringing the protest wave to an end.

[41] Roberts 2014.

https://doi.org/10.1017/S0305741015001708 Published online by Cambridge University Press

摘要: 中日钓鱼岛 (尖阁列岛) 冲突在2012年八九月间突然升级。在中国大陆既发生了街头示威， 又在新浪微博上出现了公众情绪的大爆发。我们通过人工和电脑辅助的内容分析方法利用微博原始数据来衡量整个争端中中国公众情绪的波动。我们提出两个问题: 第一, 网络民族主义情绪的凝聚力和变动性如何? 第二, 我们衡量中国政府对微博的审查, 从而进一步提问:哪些情绪在什么时候会被政府允许表达? 我们有两项主要发现: 第一, 许多微博发布者最尖锐的批评不是针对日本, 而是针对本国政府; 第二, 虽然政府在争端过程中绝大多数时间对于不同话题的审查都极为严格， 但8月18日审查力度突然直线下跌， 而这一天微博发布者针对北京政权的愤怒也正好达到顶峰。以上观察验证了以下三种理论解释， 包括两种工具性和战略性理论 (观众成本理论和减压阀理论) 和一种观念性理论 (精英对于抗议者诉求的认同)。

关键词: 社会化媒体; 新浪微博; 民族主义; 网络审查; 钓鱼岛; 日本

References

Bamman, D., B. O'Connor and N. Smith. 2012. "Censorship and deletion practices in Chinese social media." *First Monday* 17, 3–5.

Callahan, William. 2009. "The cartography of national humiliation and the emergence of China's geobody." *Public Culture* 21(1), 141–173.

Carlson, Allen. 2009. "A flawed perspective: the limitations inherent within the study of Chinese nationalism." *Nations and Nationalism* 15(1), 20–35.

Cheng, Yinhong. 2011. "From campus racism to cyber racism: discourse of race and Chinese nationalism." *The China Quarterly* 207, 561–579.

Fu, King Wa, Chung-Hong Chan and Michael Chau. 2013. "Assessing censorship on microblogs in China: discriminatory keyword analysis and the real-name registration policy." *EEE Internet Computing* 17(3), 42–50, http://doi.ieeecomputersociety.org/10.1109/MIC.2013.28.

Gries, Peter Hays. 2004. *China's New Nationalism: Pride, Politics, and Diplomacy*. Berkeley, CA: University of California Press.

Gries, Peter Hays, Qingmin Zhang, H. Michael Crowson and Huajian Cai. 2011. "Patriotism, nationalism, and China's US policy: structures and consequences of Chinese national identity." *The China Quarterly* 205, 1–17.

Hanley, James A., and Abby Lippman-Hand. 1983. "If nothing goes wrong, is everything all right?" *Jama* 249(13), 1743–45.

Hassid, Jonathan. 2012. "Safety valve or pressure cooker? Blogs in Chinese political life." *Journal of Communication* 62, 212–230.

He, Yinan. 2007. "History, Chinese nationalism, and the emerging Sino-Japanese conflict." *Journal of Contemporary China* 16(50), 1–24.

Hoffman, Robert, and Jeremy Lerner. 2013. "The demography of Chinese nationalism: a field-experimental approach." *The China Quarterly* 213, 189–204.

Hopkins, Daniel J., and Gary King. 2010. "A method of automated nonparametric content analysis for social science." *American Journal of Political Science* 54(1), 229–247.

Hughes, Christopher. 2006. *Chinese Nationalism in the Global Era*. Abingdon: Routledge.

Hussain, Muzammil M., and Philip N. Howard. 2013. "What best explains successful protest cascades? ICTs and the fuzzy causes of the Arab Spring." *International Studies Review* 15, 48–66.

Johnston, Alastair Iain, and Daniela Stockmann. 2007. "Chinese attitudes toward the United States and Americans." In Peter Katzenstein and Robert Keohane (eds.), *Anti-Americanisms in World Politics*. Ithaca, NY: Cornell University Press.

https://doi.org/10.1017/S0305741015001708 Published online by Cambridge University Press

King, Gary, Jennifer Pan and Margaret Roberts. 2013. "How censorship in China allows government criticism but silences collective expression." *American Political Science Review* 107(2), 1–18.

King, Gary, Jennifer Pan and Margaret Roberts. 2014. "Reverse-engineering censorship in China: randomized experimentation and participant observation." *Science* 345(6199), DOI: 10.1126/science.1251722.

Link, Perry, and Xiao Qiang. 2013. "From 'fart people' to citizens." *Journal of Democracy* 24(1), 79–85.

Lorentzen, Peter. 2014. "China's strategic censorship." *American Journal of Political Science* 58(2), 402–414.

MacKinnon, Rebecca. 2012. *Consent of the Networked: The Worldwide Struggle for Internet Freedom*. New York: Basic Books.

Morozov, Evgeny. 2011. *The Net Delusion: The Dark Side of Internet Freedom*. New York: Public Affairs.

Oksenberg, Michel. 1986. "China's confident nationalism." *Foreign Affairs* 65(1), 501–523.

Reilly, James. 2011. *Strong Society, Smart State: The Rise of Public Opinion in China's Japan Policy*. New York: Columbia University Press.

Reilly, James. 2014. "A wave to worry about? Public opinion, foreign policy, and China's anti-Japan protests." *Journal of Contemporary China* 23, 197–215.

Roberts, Margaret. 2014. "Fear or friction? How censorship slows the spread of information in the digital age." Paper presented at the Association of Asian Studies Annual Meeting, Philadelphia, March 2014.

Rozman, Gilbert. 2013. "Chinese national identity and East Asian national identity gaps." In Gilbert Rozman (ed.), *National Identities and Bilateral Relations: Widening Gaps and Chinese Demonization of the United States*. Stanford, CA: Stanford University Press, 203–233.

Shen, Simon. 2007. *Redefining Nationalism in Modern China: Sino-American Relations and the Emergence of Chinese Public Opinion in the 21st Century*. Basingstoke, Hants: Palgrave Macmillan.

Weeks, Jessica. 2008. "Autocratic audience costs: regime type and signaling resolve." *International Organization* 62, 35–64.

Weiss, Jessica Chen. 2013. "Authoritarian signalling, mass audiences, and nationalist protest in China." *International Organization* 67, 1–35.

Whiting, Allen. 1983 "Assertive nationalism in Chinese foreign policy." *Asian Survey* 23(8), 913–933.

Zhao, Suisheng. 2004. *A Nation-state by Construction: Dynamics of Modern Chinese Nationalism*. Redwood City, CA: Stanford University Press.

Zhao, Suisheng. 2013. "Foreign policy implications of Chinese nationalism revisited: the strident turn." *Journal of Contemporary China* 22(82), 535–553.

Zhu, Tao, David Phipps, Adam Pridgen, Jedidiah R. Crandall, and David S. Wallach. 2013. "The velocity of censorship: high-fidelity detection of microblog post deletions." eprint arXiv:1303.0597. http://arxiv.org/abs/1303.0597. Accessed 23 March 2014.

Appendix A: Procedure, Inter-coder Reliability and Keyword Validity

Three coders each independently read and scored 479 Weibo posts.[42] Each post contained any original text, plus any reposted or "retweeted" content. To

[42] This section describes the specifics of our procedure for coding the sample containing the *diaoyudao* keyword. With respect to the smaller sample of dispute-relevant posts not containing this keyword, the same basic procedure of reading the entire post text and assigning it to one of the eight categories was followed except that, owing to resource limitations, only one of the authors undertook coding. This author, although a single individual, had already benefited from several previous rounds of coding and team

simplify analysis, we counted both the original text and the reposted text (if any) as part of the same message unit, i.e. we read the posts with an eye to gauging the sentiment of this overall combination, rather than considering original and reposted sentiments separately. In Appendix Table 1, we report common reliability statistics.

Appendix Table 1: **Inter-coder Reliability Statistics**

Number of coders	3
Number of observations	479
Number of coding decisions	1,437
Average pair-wise per cent agreement	60.7
Pair-wise agreement coders 1 & 3	0.59
Pair-wise agreement coders 1 & 2	0.63
Pair-wise agreement coders 2 & 3	0.60
Fleiss' Kappa	0.53
Fleiss' Kappa observed agreement	0.61
Fleiss' Kappa expected agreement	0.17
Proportion decided by coin flip (after failed attempt to reach consensus)	.06

Additionally, we calculated a unique statistic to take account of how often we had to resort to flipping a coin to break an impasse over two codings.[43] Given the difficulty of the coding exercise, about 40 per cent of the time we resorted to brief discussion in order to reconcile different codings. These discussions usually lasted only a minute or two, and frequently one or more coders was eager to change his or her mind, having felt that he/she had incorrectly assigned a post owing to error or fatigue. Instances where coders disagreed with each other to the point where arriving at a consensus score was impossible were infrequent, and occurred only 6 per cent of the time.

Keyword validity

After completing coding, we wished to evaluate the correspondence between our human-derived sentiment categories and keyword proxies. One measure of this is what percentage of posts containing a given keyword ended up belonging to the "appropriate" category for that keyword.[44] This measure is in Appendix Table 2.

footnote continued

discussion. Therefore, although the results are more subject to his personal biases than the team results, we are confident that they are in the neighbourhood of figures that the team would have achieved.

[43] In situations where the three coders each assigned three separate scores to a post, and remained at deadlock after discussion per the above rules, we flipped a coin twice. This situation was rare and only occurred a few times.

[44] Benchmarking our keyword measures' ability to proxy for underlying categories in this manner is analogous to the "precision" measure in the computer science literature – we are more concerned about false

https://doi.org/10.1017/S0305741015001708 Published online by Cambridge University Press

The results show that the incidence of *tianchao* perfectly predicts a post belonging to the anti-government category, and the keywords *za* and *lixing* proxy moderately well for the moderate category.

Appendix Table 2: **Percentage of Posts with a Given Keyword Belonging to "Correct" Category**

Term	Correct category	Posts with (category + keyword)/posts with keyword	Percentage of posts
Anti-government (*tianchao* 天朝)	5	9/9	100
"Boycott Japanese goods" (*dizhi Rihuo* 抵制日货)	4	9/19	47
"The Diaoyu Islands are China's" (*diaoyudao shi Zhongguode* 钓鱼岛是中国的)	2	23/48	48
"Smash" (Japanese cars, etc.) (*za* (*che*) 砸 (车))	1	18/26	69
"Rational patriotism" (*lixing aiguo* 理性 爱国)	1	6/8	75
Total		65/110	59

Appendix B: Modelling the Censorship Data-generating Process

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 our sample of around 43,000 individuals, a fraction decide to write a post in response to some event.[45] Also assume that individuals who choose to write a post do so immediately following the event.[46] What we want to know is how many of these posts will survive (not

footnote continued

positives than about keywords' ability to retrieve all relevant content for a category. We are aware that by using keywords we cannot infer fluctuations in sentiment categories from changes in keyword counts over time. However, as a qualitative as well as quantitative illustration of the sorts of sentiments prevalent during the dispute, we believe our keyword approach to be a valuable complement to the directly estimated category proportions, as well as the *ReadMe* results.

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

[46] 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.

https://doi.org/10.1017/S0305741015001708 Published online by Cambridge University Press

be censored) long enough to appear in the *WeiboScope* data. This information is necessary to calculate our primary quantity of interest – the true censorship rate:

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

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 by Fu, Chan and Chau, involved periodically returning to the pages of the 43,000 users, downloading a copy of the timeline each time.[47] If a post was 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, owing to the limits set by Sina.com, the team could only crawl most of these pages (38,000 out of the 43,000, which 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. find that most censorship occurs within an hour or so of the post time,[48] 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 we have is the observed rate, R_{obs} , in Equation 2:

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

Since C_{hid} is missing, $R_{obs} < R_{true}$, i.e. the observed censorship rate is biased downward. But how much so? The *observed* rate within this project for posts with the *diaoyudao* keyword is around 12 per cent, an oddly low figure for an episode that saw the largest street demonstrations in mainland China since Tiananmen. To calculate the true rate, we need to know the true number of posts censored, N_{true}, which is related to N_{obs}, the number of censored posts that we actually observe, via some probability distribution that models the speed with which censors remove posts during episodes of collective action.[49]

Since we do not know the true distribution, we need to look for an empirical example that provides a good approximation. The best available so far is the

[47] Fu, Chen and Chau 2013.

[48] Zhu et al. 2013.

[49] An earlier version of this paper contained a mathematical formalization which is omitted here for space purposes.

https://doi.org/10.1017/S0305741015001708 Published online by Cambridge University Press

finding by Zhu et al. that "nearly 90% of deletion events happen within the first 24 hours."[50] Conveniently, this time window is the same as that of the unbiased portion of our data: 100 per cent of posts will be observed, and correctly identified as censored or not, if they survive 24 hours or more. Looking at the *WeiboScope* data for the keyword *diaoyudao* during our study period, we find that nearly all (16,967 out of 17,640) posts marked censored, or 96.2 per cent, are marked as having been deleted more than 24 hours after their posting time. For simplicity's sake, we exclude from our calculations the other 3.8 per cent of posts in the data.[51] Instead, we pragmatically treat the 16,967 posts as if they constituted *WeiboScope's* entire pool of censored posts. Since Zhu et al. found that 90 per cent of censorship occurs before 24 hours, 10 per cent must occur after, sometimes days or weeks later. Since we observe this 10 per cent, and critically, *assuming that the form of the censorship distribution over time is the same in our data as in that of* Zhu *et al.*, the ratio of what we 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 us close to the true rate. Plugging this into Equation 1 gives:

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

Applying this equation to the present study gives Appendix Table 3, which shows the observed censorship rate, the number of observed posts (including non-censored posts), and the estimated true rate. We calculate these numbers for the *diaoyudao* keyword sample, and the non-keyword sample (for key dates), and then estimate the joint rate among all topic-relevant posts.

Given that we are applying another Weibo study's findings to a different dataset, the question might arise, given that our data consist of journalists, dissidents, and verified users with more than 10,000 followers – all sensitive groups in censors' eyes – whether 90 per cent within 24 hours is too slow a rate for our sample. Zhu et al. and King, Pan and Roberts both find that a small number of ultimately censored posts typically linger for days after an incident – the question here is how many? Our main empirical concern in this paper is underestimating, not

[50] Zhu et al. 2013, 1.

[51] All censored posts in the data have a "deleted last seen" time, which marks the time at which they were last crawled before later being identified as having gone missing. Because the "deleted last seen" time for 673 of these 17,640 posts (3.8%) is less than 24 hours after the original post time, we cannot know whether these posts survived 24 hours, or indeed exactly how long they survived after being crawled (consider, hypothetically, a post that was crawled 23 hours 59 minutes after being posted, and then instantly deleted – for this post, all we know is that it survived less than 47 hours 59 minutes, or the post would have been crawled again and marked with a later timestamp). Because we do not know how many of these 673 posts survived at least 24 hours, or precisely how long they survived, we cannot use this information to "adjust" Zhu et al.'s observed distribution for our unique data; if we could, it would improve the accuracy of our adjustment formula for calculating the true rate. Given that almost all posts marked as censored in the dataset have a "deleted last seen" timestamp greater than 24 hours, however, we pragmatically choose to disregard these other 673 posts as negligible, and to exclude them from our calculations. That is, we multiply the raw counts of censored posts by 16,967/17,640 and use this number as $\mathrm{C_{obs}}$.

https://doi.org/10.1017/S0305741015001708 Published online by Cambridge University Press

https://doi.org/10.1017/S0305741015001708 Published online by Cambridge University Press

Appendix Table 3: **Observed versus True Censorship Rates for Select Dates**

Date	Posts (with keyword)	Estimated posts (w/o keyword)	Obs. rate (*diaoyudao* keyword)	True rate (*diaoyudao* keyword)	True rate for posts w/o keyword (95% CI)	Rate for whole population (95% CI)
15 Aug	6,173		9.8	52.0		
16 Aug	6,427		8.8	49.0		
17 Aug	3,581		8.4	47.9		
18 Aug	7,044	23,137	3.2	24.7	0–38.5	6.1–36.1
19 Aug	7,180		7.2	43.7		
20 Aug	2,411		11.4	56.2		
9 Sep	680		11.2	55.7		
10 Sep	3,565		14.0	62.0		
11 Sep	10,130		14.2	62.3		
12 Sep	8,480		13.4	60.8		
13 Sep	7,053		12.7	59.3		
14 Sep	9,455		13.0	59.8		
15 Sep	7,255	24,701	13.3	60.5	51.0–82.5	53.0–79.5
16 Sep	6,243		12.6	58.9		
17 Sep	5,873		13.4	60.7		
18 Sep	7,394		12.6	59.0		

over-estimating, the censorship rate. If we assume that the true number is 95 per cent within 24 hours,[52] i.e. $C_{hid} = 19C_{obs}$, then plugging these numbers into Equation 1 yields the information in Appendix Table 4. The rates given in Appendix Table 4 are higher than the previous estimates. However, the estimated mean censorship rate (with the *diaoyudao* keyword) for 18 August still falls short of 40 per cent, a rate far less than that for September and lower than that for other collective events.

[52] A full defence of this assumption is beyond the paper's scope, but here we briefly describe our logic. We think of 95% as a very conservative upper bound according to the following: if the true amount within 24 hours were indeed 95%, this would imply that a very large volume of Diaoyu-relevant Weibo content was created by users, and then wiped out of existence before being captured in the dataset. Comparing this potential volume with post surges from Weibo's top topic in 2012 (the London Olympics), we set a "face validity" limit on how large the pool of deleted posts could have been, and therefore an upper limit to the maximum percentage deleted within 24 hours. For example, we count 47,821 as the number of posts in *WeiboScope* containing the keyword "Olympic" (*aoyun* or *aolinpike*) on 28 August 2012, the date on which the opening ceremony for the London Olympics was broadcast on Beijing time. Then, using the above figure, we assume that this keyword was heavily censored as if it were a proxy for a collective action topic (a dubious, worst-case assumption, given that most discussion about the Olympics was surely non-political), and we use Zhu et al.'s 90% estimate in extrapolating and "adding back" a large hypothetical number of censored posts. The phrase "2012 London Olympics" (2012 *nian lundun aoyun hui* 2012) was Weibo's top trending topic of 2012 according to Sina.com; in comparison, the phrase "the Diaoyu Islands are China's" (*diaoyudao shi Zhongguode*) ranked tenth. If we allow that the total number of *pre-censorship* Diaoyu-relevant posts on 18 August could not have been greater than the Olympics-related figure above, e.g. Diaoyu posts<47,821, then for this inequality to hold, the percentage within 24 hours could not have exceeded about 91.45%. Given this, we think that an estimate of 95% is exceedingly high and well beyond a more feasible maximum; we choose this high number to demonstrate the robustness of our results subject to all assumptions presented here.

https://doi.org/10.1017/S0305741015001708 Published online by Cambridge University Press

https://doi.org/10.1017/S0305741015001708 Published online by Cambridge University Press

Appendix Table 4: **Observed versus True Censorship Rates under 95 per cent Assumption**

Date	Posts (with keyword)	Estimated posts (w/o keyword)	Obs. rate (*diaoyudao* keyword)	True rate (*diaoyudao* keyword)	True rate for posts w/o keyword (95% CI)	Rate for whole population (95% CI)
15 Aug	6,173		9.8	68.4		
16 Aug	6.427		8.8	65.8		
17 Aug	3,581		8.4	64.8		
18 Aug	7,044	23,137	3.2	39.6	0–55.6	11.6–53.1
19 Aug	7,180		7.2	60.8		
20 Aug	2,411		11.4	72.0		
9 Sep	680		11.2	71.6		
10 Sep	3,565		14.0	76.5		
11 Sep	10,130		14.2	76.8		
12 Sep	8,480		13.4	75.6		
13 Sep	7,053		12.7	74.5		
14 Sep	9,455		13.0	74.9		
15 Sep	7,255	24,701	13.3	75.4	67.6–90.4	69.3–88.6
16 Sep	6,243		12.6	74.2		
17 Sep	5,873		13.4	75.6		
18 Sep	7,394		12.6	74.2		