Tweeting Under Pressure: Analyzing Trending Topics and Evolving Word Choice on Sina Weibo

Le Chen
College of Computer and Information Science
Northeastern University
Boston, MA USA
leonchen@ccs.neu.edu

Chi Zhang
College of Computer and Information Science
Northeastern University
Boston, MA USA
czhang79@ccs.neu.edu

Christo Wilson
College of Computer and Information Science
Northeastern University
Boston, MA USA
cbw@ccs.neu.edu

ABSTRACT

In recent years, social media has risen to prominence in China, with sites like Sina Weibo and Renren each boasting hundreds of millions of users. Social media in China plays a profound role as a platform for breaking news and political commentary that is not available in the state-sanctioned news media. However, like all websites in China, Chinese social media is subject to censorship. Although several studies have identified censorship on Weibo and Chinese blogs, to date no studies have examined the overall impact of censorship on discourse in social media.

In this study, we examine how censorship impacts discussions on Weibo, and how users adapt to avoid censorship. We gather tweets and comments from 280K politically active Weibo users for 44 days and use NLP techniques to identify trending topics. We observe that the magnitude of censorship varies dramatically across topics, with 82% of tweets in some topics being censored. However, we find that censorship of a topic correlates with high user engagement, suggesting that censorship does not stifle discussion of sensitive topics. Furthermore, we find that users adopt variants of words (known as *morphs*) to avoid keyword-based censorship. We analyze emergent morphs to learn how they are adopted and spread by the Weibo user community.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and behavioral sciences; K.5.2 [**Governmental Issues**]: Censorship

Keywords

Online social networks; Sina Weibo; Trending topics

1. INTRODUCTION

In recent years, social media has risen to prominence in China. Sina Weibo (the Chinese equivalent of Twitter, abbreviated as *Weibo*) boasts 500 million users [45], and Renren (the Chinese equivalent of Facebook) boasts 172 million users [22]. Like people the world over, Chinese users flock to these platforms as places to socialize and share content. However, social media in China also plays a more profound role as a platform for breaking news and political commentary that is not available in the state-sanctioned news media. For example, Weibo played a key role in the downfall of once-prominent politician Bo Xilai [17].

Like all websites in China, Chinese social media is subject to government-enforced content regulation policies. The primary manifestation of these regulations is censorship, which is known to impact Chinese blogs [25] and Weibo. Current work disagrees on the scope of censorship on Weibo, with estimates ranging from 0.01% [42] to 16% [7] of all "weibos" (*a.k.a.* tweets on Weibo) being censored. Users who discuss political issues [42, 49] and minority groups [7] tend to incur the brunt of censorship. In fact, it is hypothesized that Weibo employs thousands of crowdsourced workers to manually examine and censor the huge volume of tweets that are generated each day [49]. Thus, tweets may be visible for minutes, hours, or even days before they are censored, giving researchers an opportunity to download and analyze them.

Although it is no secret that tweets on Weibo are censored, how censorship is applied and the impact that it has on discourse is currently unknown. In this study, we seek to answer two key questions: *first*, what is the impact of censorship on discourse on Weibo? In other words, is censorship effective at chilling or even halting discussion on Weibo? *Second*, do Weibo users adapt in order to avoid censorship? Anecdotal evidence suggests that users may use *morphs* to avoid keyword-based censorship [7, 42], *e.g.,* 储君 (crown prince) instead of 习近平 (Xi Jinping, the current president of China). However, it is unknown whether this theory is true, and if so, what the dynamics of morph generation are. These two questions get at the heart of the conflict between information dissemination and censorship in the highly dynamic, human-driven social media space.

To answer these questions, we break our study down into three major components. *First*, we conduct a large scale crawl of Weibo for 44 days. Our crawl targeted a connected component of 280,250 users who are active on Weibo. The crawler implemented a prioritization system where users who tweet more frequently were crawled more frequently. This enabled the crawler to gather most censored tweets before they were deleted (censorship can then be identified after-the-fact). In total, our crawl gathered 36.5M tweets, 1% of which were censored. We observe that censorship is not applied uniformly, *e.g.,* 82% of tweets from one particularly contentious topic were censored, while up to 50% of tweets from some celebrity users were censored.

In addition to tweets, our crawler also gathered all of the *comments* on each tweet. Comments on Weibo function like comments on Facebook, *i.e.,* users append them to existing tweets. Unlike prior studies of censorship on Weibo, ours is the first to examine

COSN'13, October 7–8, 2013, Boston, Massachusetts, USA.

http://dx.doi.org/10.1145/2512938.2512940.

both tweets and comments. This distinction is important, because, as we show in § 4.2, there are an order of magnitude more comments than tweets on Weibo.

Second, we leverage Latent Dirichlet Allocation (LDA) [9] to extract 37 trending topics from our crawled data. Each one of these topics corresponds to a real-world event (*e.g.,* the Boston marathon bombing, Ya'an Earthquake, *etc.*), and several were heavily censored (*e.g.,* a Sichuan official who was criticized following the Ya'an earthquake, an incident between President Xi and a Beijing taxi driver, *etc.*). Across these topics, we analyze the relationship between the magnitude of censorship and the characteristics exhibited by the topic (*e.g.,* number of engaged users, tweets per user, *etc.*). Contrary to our expectations, we find that users are more active in discussing censored topics, indicating that censorship does not have a chilling effect on discussion on Weibo.

Third and finally, we examine the usage of morphs on Weibo. We find that 11 of our 37 topics include morphs, in some cases up to 5 morphs per topic. Although we observe that many uncensored topics include morphs for comedic or satirical effect (*e.g.,* 黑十字 (Black Cross) in place of 红十字 (Red Cross)), we also find that morph usage dramatically increases within censored topics. Temporal analysis reveals that morph usage increases rapidly within hours of censorship being implemented, suggesting that users adapt their word usage to circumvent censorship.

We view this study as a first step towards understanding the impact of censorship on discourse in social media, rather than simply quantifying the scope of censorship. This study lays the foundation for updating existing information dissemination models, or developing new ones, that take adversarial forces into account. Our results also point towards new techniques for identifying and predicting censorship, by using language models to observe when words usage changes (*i.e.,* morphs) in otherwise unexpected ways.

2. BACKGROUND

We begin by briefly introducing Sina Weibo, comparing its features to Twitter, and discussing government regulation of the Web in China.

2.1 Sina Weibo

Sina Weibo (referred to as Weibo) is the most popular microblogging website in China. Weibo first launched in August 2009 and by December 2012, it had ≈500 million users. Over 4.6 million users are active on a daily basis, and over 100 million weibos (*a.k.a.* a tweet on Weibo) are posted every day [44, 45]. As of April 2013, Alexa shows that Weibo is the No.6 website in China, and No.29 website globally.

Weibo provides similar functionality to Twitter. Users can *follow* other users and view their tweets in a timeline. Users post 140-Unicode-character tweets which can include URLs, pictures, videos, geotags, *retweets*, $@mentions$, and $\#hashtags$. Each Weibo user has a personal profile that may include basic information (*e.g.,* hobbies, hometown, *etc.*) as well as statistics (*e.g.,* total tweets, followers, and followings). Like Twitter, Weibo users can be "verified," *i.e.,* manually vetted by Weibo staff to confirm their identity.

Weibo also offers some features that are similar to Facebook. Weibo users may *like* and/or attach *comments* to tweets. Weibo users may also prepend 140-character messages to retweets. The ability to comment gives conversations on Weibo a well-defined, multi-layered structure. Comments may contain $@mentions$ and $\#hashtags$, just like tweets.

Like Twitter, Weibo provides rate-limited APIs to developers. These APIs enable software to retrieve users' timelines, post and delete tweets, *etc.* However, as we discuss in § 3.1, there are significant limitations to Weibo's APIs.

2.2 Government Regulation of the Web

The Chinese government enforces several policies to regulate content on websites. As a major social hub, Sina Weibo regulates content in cooperation with these policies.

- **Real Name Policy.** In March 2012, Weibo implemented the Real Name Registration (RnR) Policy [48]. The policy states that users must use their real name when creating a Weibo account, although users may use a pseudonym as their public handle on the website.
- **Blacklists.** Weibo maintains a blacklist of words and URLs that are not permitted in tweets. For example, tweets may not contain links that leverage Google's URL shortener `goo.gl`.
- **Search Censorship.** Weibo does not permit users to search for tweets that contain certain words. The China Digital Times maintains an up-to-date list of words impacted by search censorship on Weibo [1].
- **Tweet Censorship.** Several studies have confirmed that Weibo censors tweets [7, 49]. Tweets may be deleted if they contain politically sensitive topics, abusive language, pornography, or rumors. It is hypothesized that Weibo employs a heterogeneous strategy for tweet censorship, ranging from keyword filtering to real-time crowdsourced monitoring [49].

Violations and Penalties. Sina Weibo enforces penalty policies against users who violate online content regulations. The *Sina Weibo Community Treaty*, launched in May 2013 [6], outlines these penalty policies. The treaty introduced a *credit* system for Weibo users where *credit* is deducted for each policy violation. Weibo accounts are permanently deleted if their *credit* reaches 0. Accounts may also be temporarily suspended at Sina Weibo's discretion. Many violations are detected and handled by Weibo's automatic security systems, *e.g.,* spam tweets, tweets that link to pornography, and tweets that include blacklisted keywords. More complex violations (such as disseminating false or misleading rumors) are handled by Weibo's *Community Board*, which is composed of well-known Weibo users that are hand-chosen by Sina.

Awareness of and Responses to Censorship. Sina Weibo users are aware that the social network censors content. Users must agree to the *Sina Weibo Community Treaty* when they register for an account, and complaints about censorship are common on Weibo, as well as on other Chinese web forums. Although the *Treaty* does not specify what topics or words are censored, there are webpages that catalogue the details of censorship on Weibo [1]. Thus, savvy web users can locate the current list of censored topics and words on Weibo.

Given this awareness of censorship, Chinese Web users have adopted a variety of obfuscation techniques to avoid censorship. In particular, users have been observed using abbreviations, anglicanizations of Chinese characters, neologisms (newly invented words), homophones (words that sound the same), and homographs (words that look similar) to avoid keyword-based censorship [25, 49]. Collectively, we refer to these words as *morphs*.

Although we cannot be certain that any given Weibo user is aware of censorship, as we show in § 6, users adopt morphs in tandem with the emergence of politically sensitive trending topics.

This morph adoption begins even before censorship is imposed, indicating that, in general, users are aware of censorship and try to avoid it preemptively.

2.3 Studies of Censorship on Weibo

Three existing studies examine censorship on Weibo. Bamman *et al.* confirmed the existence of censorship by calculating that tweets with certain words are deleted much more frequently than predicted by random chance [7]. Fu *et al.* developed statistical tools to locate censored keywords, and examined the chilling effect of RnR on Weibo [42]. However, these studies disagree on the scope of censorship on Weibo, with the former claiming that 16% of tweets are censored, and the latter claiming 0.01% of tweets are censored. These drastically different estimates may be due to different tweet sampling methodologies between the two studies. Finally, Zhu *et al.* measure the velocity of censorship, and observe that 30% of censored original tweets (*i.e.,* not retweets) are deleted within 30 minutes [49].

Censorship of Comments on Weibo. Although existing studies confirm that Weibo censors tweets, to date no studies have examined censorship of comments on Weibo. During our study, we observed that when a tweet is deleted (for any reason), the comments associated with that tweet are also deleted. We also observed Weibo deleting comments that contain malicious links and spam. However, *we have not observed any instances where Weibo has censored a comment*. We confirmed this observation by searching the Chinese Web for complaints about censorship on Weibo: although many users complain about tweet censorship, we could not find a single instance of users complaining about comment censorship. This distinction is important, because, as we show in § 4, there are an order of magnitude more comments than tweets on Weibo.

3. METHODOLOGY

The goal of this study is to examine the impact of censorship on topical discussion and word usage on Weibo. In particular, we want to address two broad questions: *first*, is censorship effective at diminishing (or even halting) discussion of particular topics? *Second*, do Weibo users adapt to try and avoid censorship (*e.g.,* by using morphs), and if so, what are the dynamics of this process?

To answer these questions, we need to collect a large corpus of tweets and comments from Weibo over a long period of time. In this section, we present our methodology for gathering this data. First, we discuss the challenges presented by collecting data from Weibo. Next, we introduce the population of users targeted by our crawler. Finally, we discuss the design of our prioritized crawler, and validate its effectiveness at gathering censored tweets.

3.1 Data Gathering: API or DIY?

There are three options for gathering data from Weibo: sampling tweets from the public timeline API, querying the tweets of individual users with the developer API, or crawling the website. We chose to crawl the Weibo site for two reasons. First, Weibo's public timeline API (which is roughly equivalent to Twitter's "spritzer" data stream) does not include retweets or comments. As we show in § 4, retweets and comments account for 97% of the content on Weibo. Thus, the public tweet API is unusable for our study.

Second, Weibo's developer APIs are inefficient for gathering tweets and comments. Each call to the API returns the most recent 100 tweets for a given user, however an additional API call is necessary to gather the comments *on each tweet*. In contrast, the Weibo site return 10 tweets per HTTP request, along with the first 10 comments on each of those tweets. As we shown in § 4, >99% of tweets accrue ≤10 comments, meaning that 10 HTTP requests is roughly equivalent to making 101 API calls.

3.2 Selecting Weibo Users

The next step in our study is identifying a subset of Weibo users to crawl. We choose to focus on a large, diverse, connected component of users rather than a random sample because studies have shown that different types of Weibo users experience dramatically different levels of censorship. For example, Fu *et al.* find that 0.01% of tweets from 350K celebrities (users with >1000 followers) are censored [42], whereas Zhu *et al.* find that 13% of tweets from 3K politically active Weibo users are censored [49].

Seed Selection. To locate a connected component of Weibo users, we first select 7 politically active Chinese celebrities as *seeds*. We then gathered all of the users who the seeds follow, most of whom are also celebrities in China. Collectively, we refer to these 3049 users as *celebrities*.

Selecting Commentors. The next step is to add normal users to the connected component. Unfortunately, it is not feasible to crawl the 33M followers of the *seeds* on a daily basis. Furthermore, it has been shown that 57% of Weibo accounts never tweet, and 90% tweet less than once per week [41]. Thus, randomly selecting from the *celebrity* followers is unlikely to uncover active users.

Instead, we select normal users from the set of users who comment on tweets from the seeds. We crawled all comments on all tweets from the 7 seeds between October 2012 and February 2013. This process located 2.8M commentors, which is still too many to crawl on a daily basis. We decided to split our resources by crawling the 177K *top commentors* and a 100K sample of *random commentors* (note that these two populations are non-overlapping). Thus, our final target population includes 280,250 users.

This split between *top* and *random* commentors allows us to crawl highly active users, as well as a less biased sample of average users. Each *top commentor* generated ≥10 comments during the measurement period, while ≈60% of the *random commentors* only commented once. We observe that *commentors* who comment more than once tend to do so at two week intervals. We compare the characteristics of our three target groups in more detail in § 4.

3.3 Crawler Design and Data Collection

Now that we have selected the target population, we must develop a strategy to crawl these users. On one hand, we want to crawl each user's tweets as often as possible, since it has been shown that censored tweets can be deleted in a matter of minutes [49]. On the other hand, the number of HTTP requests we can make to Weibo each day is rate-limited, and we want to collect data from a large number of users.

Prioritized Crawler Design. To balance these competing goals, we develop a prioritized crawler. Prior work has shown that Weibo users tweet at different rates [41]. This indicates that our crawler should visit some users more frequently than others.

To correctly allocate our resources, we need to understand the tweeting behavior of the target user population. Thus, we crawled all the tweets generated by the target users in January 2013. Figure 1 shows the inter-arrival time between tweets for *celebrities*, *top commentors*, and *random commentors*. The figure shows CDFs of minimum, average, and maximum inter-arrival times for all users. Two conclusions can be drawn from Figure 1. First, all three user populations have similar overall behavior. Second, the vast majority of users tweet between once every three hours, and once per day. A small fraction of users tweet more frequently.

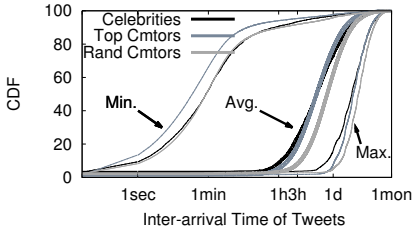

Figure 1: Inter-arrival time between tweets for our three target groups.

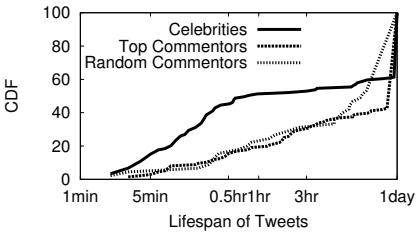

Figure 2: Lifespan of censored tweets.

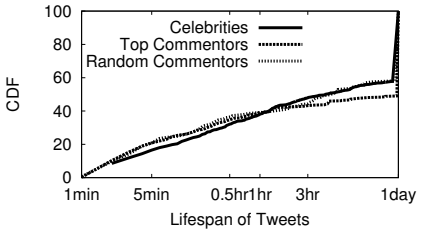

Figure 3: Lifespan of tweets deleted by their owner.

Based on the results in Figure 1, we can implement a prioritized crawler. The crawler has three buckets: one hour, three hours, and daily. A user in a given bucket is visited by the crawler at the corresponding frequency. The buckets contain 5K, 22K, and 253K users, respectively. Most users in the one and three hour buckets are *celebrities* and *top commentors*.

Data Collection. We crawled Weibo from March 30 to May 13, 2013. The crawler collected tweets from users' timelines between 8 a.m. and 2 a.m. China Standard Time. Between 2 a.m. and 8 a.m. (when users are likely to be asleep) the crawler went back and downloaded the comments attached to all tweets found the previous day. This enabled tweets to accrue comments for many hours before we collected them. We record the unique ID, content, author, and timestamp of each tweet and comment. In addition to the prioritized, targeted crawl, we also collected 700K tweets per day from Weibo public timeline API.

Identifying Censored Tweets. We conducted periodic crawls to identify censored tweets. Every two weeks, a separate crawl would revisit each targeted user's timeline and compare the contents with our historical records. Any missing tweets would be individually queried to determine if it was censored, marked as spam, or deleted by the owner. Weibo returns explicit error messages describing why tweets are deleted, enabling researchers to unambiguously identify censored tweets. Prior work also leverages this methodology to identify censored tweets [7, 42, 49].

Dealing with Spam. Weibo is now a popular target for spammers, just like Twitter [18, 8, 38, 39]. For this study, we adopted a best-effort approach to eliminating spam from our dataset. Before finalizing the set of *commentors*, we filtered out all users with obviously suspicious interaction patterns, *e.g.,* a huge amount of tweets from a recently created account, or many comments posted within seconds of each other. We manually inspected these suspicious accounts and confirmed that they were spammers.

Despite these precautions, 4459 (1.6%) of the *commentors* were suspended from Weibo during our study. It is not clear what violation(s) of the *Community Treaty* caused these suspensions. Fortunately, the number of suspended users is very small, and does not jeopardize the fidelity of our study.

3.4 Validation

Our crawling methodology makes an explicit tradeoff in favor of scope at the expense of timeliness. Specifically, our crawler gathers tweets and comments from a large number of users, at the cost of only being able to visit each user's timeline every few hours. However, prior work has shown that tweets can be censored in as little as a few minutes [49]. This raises an important question: what percentage of censored tweets is our crawler able to gather?

To answer this question and validate our methodology, we performed an experiment: we selected 500 random users from each of our three target populations and crawled their timelines once per minute for a week (April 29 to May 5, 2013). This high-fidelity crawl enables us to calculate the lifetime of deleted tweets down to the minute. During this week, we observed 25,735 tweets, 603 of which were censored, and 1,277 of which were deleted by their owners. Note that in this experiment, we only monitor tweets from the prior 24 hours for censorship/deletion.

Figure 2 plots the lifespan of censored tweets for the three target groups. For *celebrities*, ≈50% of censored tweets are deleted within one hour of their creation, while ≈40% are censored after one day. This result is similar to the findings of Zhu *et al.* [49]. However, for *commentors* <20% of censorship occurs within the first hour. It is not clear whether the different speeds of censorship occur because celebrities are more heavily monitored by the authorities, or because they generate more objectionable tweets than *commentors*.

To put the results in Figure 2 into perspective, we plot Figure 3, which shows the lifetime of tweets deleted by their owners. In this case, the lifespan of tweets is the same across all three populations: ≈40% of tweets are deleted within the first hour. This result makes intuitive sense: if a user wants to delete one of their own tweets (*e.g.,* it contains a typo or an incorrect link), they perform this action quickly.

Implications. The takeaway from this validation experiment is that our prioritized crawler will miss some censored tweets (*i.e.,* the tweets will be generated and censored before our crawler can observe them). In the worst case, the prioritized crawler may miss 50% of censored tweets from *celebrities*, since they are crawled once every hour. Similarly, in the worst case, the prioritized crawler will miss 20-40% of censored tweets from *commentors*.

Although our crawler will miss some censored tweets, this does not adversely impact our study for two reasons. *First*, our crawler captures the majority of censored tweets from the target users. This gives us a large enough sample to know, with high statistical confidence, which topics and words are being censored. This information is sufficient to support our goal of analyzing censorship's impact on topical discussion and word usage on Weibo. *Second*, as we show in § 4.2, the vast majority of content on Weibo is in comments, not tweets. Unlike previous studies of censorship on Weibo which ignore comments [7, 42, 49], our crawler is able to capture comments. We feel that this is a favorable tradeoff, *i.e.,* gathering 18M comments per day at the expense of missing some censored tweets.

4. GENERAL ANALYSIS

In this section, we analyze the overall characteristics of our Weibo dataset. *First*, we contrast the characteristics of *celebrities*, *top commentors*, and *random commentors* with a random sample of Weibo users taken from the public timeline API. This comparison enables us to quantify the differences between our target population

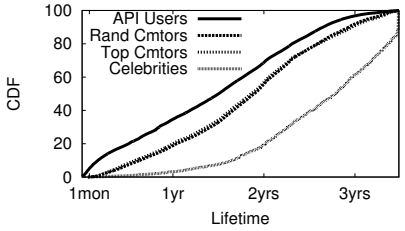

Figure 4: Lifetime of Weibo users.

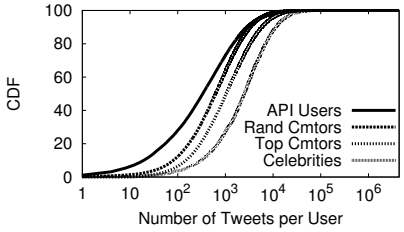

Figure 5: Total tweets per Weibo user.

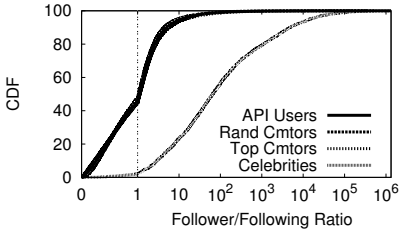

Figure 6: Follower/following ratio for Weibo users.

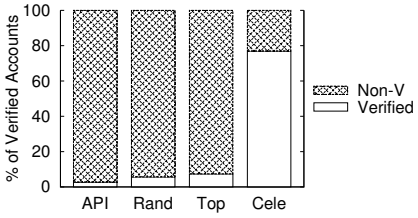

Figure 7: Percentage of verified Weibo users.

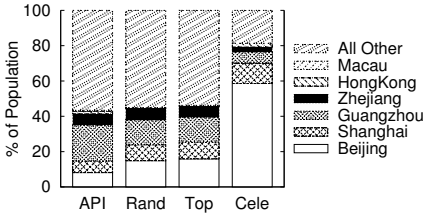

Figure 8: Location demographics for Weibo users.

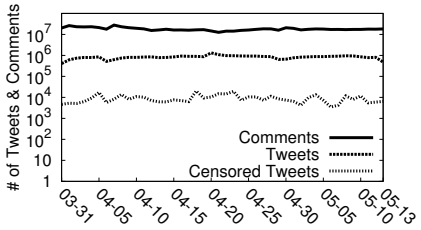

Figure 9: Tweets, comments, and censored tweets per day.

and the overall Weibo userbase. *Second*, we introduce and briefly examine the data from our daily crawls of Weibo. This sets the stage for deeper analysis of trending topics on Weibo in Section 5.

4.1 Characterizing Weibo Users

In this section, we analyze the characteristics of our three target populations by examining data from their user profiles. On Weibo, each user profile includes the date the account was created, the total number of tweets, followers, and followings for the account, whether the user is verified, and the user's self-reported geographic location. To compare the profiles of the target users with generic Weibo users, we randomly picked 1M users who appeared in the public timeline API. We refer to these users as *API users*. All profiles were crawled on February 17, 2013.

Lifetime of Accounts. First, we examine the lifetime of users in our four different groups. Since Weibo went public on August 14, 2009, the maximum user lifetime on Weibo is 1284 days. Figure 4 plots the lifetime of users in the four groups. *API users* tend to have the youngest accounts (50% are $\leq$1.5 years old). Given that the population of Weibo has been growing exponentially, it makes sense that many users have young accounts. In contrast, the two *commentor* groups have indistinguishable lifetime characteristics, and are older than the *API users* by several months. The *celebrities* have the oldest accounts (40% are $\geq$3 years old), showing that they were early adopters of Weibo.

Tweets Per User. Next, we examine how active the different user groups are by looking at the total number of tweets they generate. Figure 5 reveals that each group of users generates a different amount of tweets. The *API Users* generate the least tweets, which accords with their short lifetimes, and prior work showing that many Weibo accounts tweet infrequently [41]. *Top commentors* generate more tweets than *random commentors* despite having similar lifetimes. This corresponds to our original selection criteria, *i.e., top commentors* were chosen because they are active. *Celebrities* generate the most tweets by far because they are highly active and have the oldest accounts.

Followers vs. Followings. To gauge the impact of fame, we plot the follower/following ratio for the four user groups in Figure 6. We filter out 17K *API users*, 348 *commentors*, and 7 *celebrities* that have 0 followings. Figure 6 demonstrates that most users on Weibo have similar ratios of followers to followings, with *celebrities* being the exceptions. 98% of *celebrities* have ratios $>$1, and 77% have ratios $>$10. In contrast, 44% of *API users* and *commentors* have ratios $<$1, *i.e.,* they follow many users, but have few followers.

Verified Accounts. Similar to Twitter, Weibo provides an identity verification system for famous users (not the general public). To become verified, users must submit supporting documentation to authenticate themselves, which is then manually verified by Weibo staff.

Figure 7 plots the percentage of verified users in each of our four user groups. 77% of *celebrities* are verified, which confirms our classification of these users. $\approx$7% of *commentors* are verified, while only 3% of *API users* are verified.

Geographic Distribution. Finally, we study the geographic distribution of the four user groups. Weibo users must list a home location on their profile, with the available options being Chinese provinces, autonomous/special administrative regions, "abroad," or "other." Note that user's locations are self-reported, and may not be accurate.

Figure 8 plots the location demographics for our four user groups. Beijing, Shanghai, Guangzhou, and Zhejiang are the top 4 locations across all groups. This is not surprising, since these coastal regions all have above average rates of Internet penetration in China [32]. 59% of *celebrities* are in Beijing, possibly because it is the capital and political center of China. In contrast, the *commentors* and *API users* have similar demographics, with the former slightly favoring Beijing, and the latter Guangzhou.

Summary. The results in this section contrast our target user groups and a random sample of active Weibo users. Overall, the *commentors* and *API users* are quite similar, *e.g.,* similar follow ratios and geographic distributions. However, the *commentors* do have older accounts than *API users*. This data suggests that the *commentors*, who are 99% of our target population, are representative of active Weibo users in general.

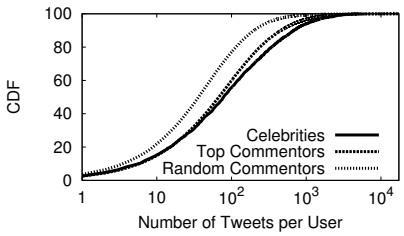

Figure 10: Tweets per user in our three target groups.

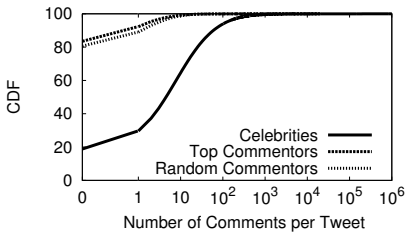

Figure 11: Comments per tweet for our three target groups.

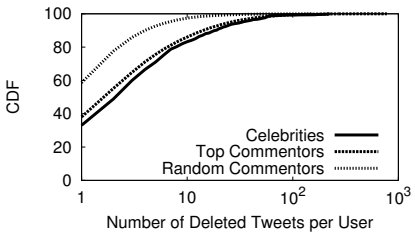

Figure 12: Censored tweets per user for our three target groups.

The *top commentor* and *random commentor* groups are extremely similar. Although we chose the *top commentors* specifically because they are very active, their overall characteristics are almost the same as the *random commentors*, who were chosen using a less biased selection process.

Unsurprisingly, the *celebrities* are very different from other Weibo users. Given that the *celebrities* only comprise 1% of our target population, these differences have little impact on the overall composition of our target population.

4.2 Daily Activity on Weibo

In this section, we provide a brief overview of the data from our daily crawls of Weibo, *i.e.,* how many tweets and comments per day, how many comments per tweet, and how many tweets are censored per day. We analyze this dataset in greater depth in § 5 and § 6.

Overall Data Collection. We conducted daily crawls of the *celebrities*, *top commentors*, and *random commentors* between March 30 and May 13, 2013. Figure 9 shows the number of tweets and comments gathered each day, along with the number of censored tweets. The number of interactions per day is roughly constant: ≈830K tweets, 18M comments, and 9K censored tweets. There are an order of magnitude more comments per day than tweets. Although it is not shown in Figure 9, there are also an order of magnitude more retweets every day than original tweets.

Figure 10 plots the number of tweets per user during our 44 days of crawled data. *Celebrities* and *top commentors* have almost identical behavior over this time period, with ≈40% of users tweeting >100 times. In contrast, the majority of *random commentors* tweet <36 times.

Figure 11 shows the number of comments attached to each tweet in our dataset. Despite the fact that *top commentors* generate more tweets than *random commentors*, both groups accrue similar amounts of comments: ≈80% of tweets receive 0 comments, and <1% receive >10 comments. In contrast, 50% of *celebrity* tweets accrue >5 comments. Clearly, *celebrity* tweets serve as hubs of discussion on Weibo.

Censorship per User. During our crawl, we observed that 1% of tweets are censored every day. However, this is a conservative estimate, given that our crawler is expected to miss some censored tweets (see § 3.4). Figure 12 plots the number of censored tweets per user in each user group. For *celebrities* and *top commentors*, ≈35% have 1 censored tweet, while ≈17% have ≥10 censored tweets. In contrast, 59% of *random commentors* only have 1 censored tweet.

5. TOPIC ANALYSIS

At this point, we have described our crawling methodology, and presented an overview of the users and timeline data gathered by it. We now return to the first of two major questions asked in this paper: *what is the impact of censorship on discourse on Weibo*? To answer this question, we extract trending topics from our Weibo data and examine the relationship between censorship and topic-level characteristics.

We organize this section into two parts. *First*, we present our methodology for locating tweets and comments that correspond to trending topics. *Second*, we introduce the 37 trending topics we identify on Weibo (including several censored topics) and analyze the correlation between censorship and topic-level characteristics.

5.1 Locating Trending Topics

Before we can analyze topic-level characteristics, we must develop a methodology for locating trending topics amongst the 839M tweets and comments collected by our crawler. We divide this process into four phases: word segmentation, topic extraction, validation, and labeling.

Word Segmentation. The first step in our methodology is segmenting tweets and comments from Weibo into individual words. This step is necessary because the Chinese language does not include breaks between words. However, identification of individual words is a necessary precondition for using many Natural Language Processing (NLP) algorithms to extract topics from text corpora.

We segment tweets and comments using OpenCLAS [23], which is an open-source implementation of the ICTCLAS Chinese word segmentation algorithm [47]. We chose ICTCLAS because it is consistently a top contender at the SIGHAN Chinese NLP bakeoff.[1] This tournament is the de-facto benchmark for state-of-the-art Chinese NLP techniques.

The weakness of OpenCLAS is that it relies on a dictionary of 104K traditional Chinese words to perform segmentation. This dictionary does not include any of the new words or morphs present on social media. To overcome this deficiency, we augmented the OpenCLAS dictionary with 6.1M words taken from the Sogou Pinyin dictionary [4]. Sogou Pinyin is a the most popular Chinese character input software in China [5], and the dictionary of words leveraged by the software is constantly updated by users who upload new words. We manually verified that OpenCLAS with the updated dictionary was able to correctly segment 1000 randomly selected tweets from our dataset.

Topic Extraction. The second step in our methodology is to extract topics from the corpus of segmented tweets and comments. For this task, we leverage LDA [9]. Although LDA cannot usually be applied to microblog text because each tweet is too short [35, 20, 34], two factors make LDA feasible in our case. First, Chinese text is denser than English, *i.e.,* more words fit into 140-character tweets. Second, on Weibo, many tweets have associated comments.

[1] http://www.sighan.org/

Topic name	Topic Description	Lifespan (Days)	Tweets	Cmts	Likes	RTs	% Censored
Lushan	Derision of a Chinese official with an expensive wristwatch after the Ya'an earthquake.	6	928	8K	265	10K	81.6%
Taxi	An incident involving a taxi driver who claimed to meet President Xi.	2	2K	3K	2K	70K	36.2%
Bird Flu	Rumors about the return of SARS horrors during the emergence of H7N9 bird flu.	4	394	10K	243	5K	20.0%
Jingwen	The suicide (rumored homicide) of a young woman at Jingwen shopping mall.	5	9K	141K	4K	72K	12.2%
Obama	White House petition asking for deportation of the suspected poisoner of Ling Zhu.	3	26K	640K	23K	268K	5.8%

Table 1: Top 5 topics ranked by percentage of censored tweets.

Topic Name	Original Words	Morphs
Lushan	范继跃 (the official's name)	芦山县委书记 (Lushan secretary), 表印哥 (brother watch-print), 无表哥 (brother no-watch), 机智哥 (brother wisdom)
Taxi	郭立新 (the driver's name)	北京的哥 (Beijing taxi driver), 郭师傅 (Shifu Guo)
Bird Flu	十年前非典 (SARS, 10 years ago), 十年后禽流感 (bird flu)	*No Morphs*
Jingwen	京温 (Jingwen), 袁利亚 (the girl's name), 钟涛 (Jingwen boss' name)	京wen (partial anglicanization of Jingwen), 袁莉亚 (homograph of the girl's name), 京温老总 (Jingwen boss), 安徽女子 (girl from Anhui), 袁某 (Yuan XX)
Obama	奥巴马 (Obama), 白宫 (the Whitehouse)	美国信访办 (US petition office), 信访办主任 (director of the petition office), 奥青天 (Oba-the-sky)

Table 2: Original words and their corresponding morphs amongst the top 5 censored topics.

For the purposes of topic extraction, we combine each tweet with its comments to form a single, longer *document*.

We applied LDA to a random sample of 1.4M documents from our timeline dataset. Before processing we filtered out rare words that appear $\leq$3 times, the top 500 most common words, a stoplist of emoticons and other useless words, and URLs. These filters increase the accuracy and decrease the running time of LDA. We set the number of topics $K = 300$, $\alpha = 0.167$, and $\beta = 0.001$ (based on the parameterization from [43]), and ran LDA for 1000 iterations. The output of LDA is 300 topics, each containing 100 words ranked by how strongly they correspond to that topic.

Manual Validation. The third step in our methodology is manually vetting and validating the topics from LDA. Manual analysis of the 300 topics revealed that 36 corresponded to real-world events that took place between March 30 and May 13, 2013. We refer to these 36 as *trending topics*. The remaining topics from LDA were very general, and did not correspond to any particular real-world event. Example topics include: gender specific terms, weather related terms, emoticons, and advertising related terms.

To validate whether the trending topics from LDA cover the popular topics on Weibo during the measurement period, we compared our 36 topics to two external sources. First, we compared the LDA topics to a list of known trending topics from April and May 2013 [2, 3]. All 11 topics (10 from April, 1 from May) are included in the 36 topics. Second, we compared the words in the 36 topics to the list of censored words from China Digital Times [1]. Words from one censored topic were not included in our 36 topics; we manually added this missing topic to our collection. These tests confirm that LDA captures the vast majority of trending and censored topics on Weibo. The one manually added topic brings our complete collection to 37 trending topics.

Labeling Tweet and Comments. The final step in our methodology is to label tweets and comments with their corresponding topic. To bootstrap this process, we identified between 1 and 4 keywords in each of our 37 topics that uniquely correspond to that topic. The vast majority of these keywords are proper nouns (*e.g.,* names and places), *e.g.,* 范继跃, 郭立新, and 袁利亚. In 11 topics, we also identified between 1 and 5 unique morphs that only occur within that topic, which we also use as keywords. We revisit these morphs later in § 6. For each tweet and comment in our dataset, we label it as topic t if it contains $\geq$1 of the keywords from topic t.

5.2 Analysis of Trending Topics

In this section, we analyze the 37 trending topics in our dataset. First, we briefly introduce the trending topics and present their high-level features. Second, we use Spearman's ρ to calculate the correlation between censorship and the characteristics of trending topics on Weibo.

High-Level Overview. The 37 trending topics in our dataset cover many topical events that occurred between March 30 and May 13, 2013. This includes major world events, *e.g.,* the death of Margaret Thatcher and the Boston Marathon bombing. They also cover important events within China, *e.g.,* the dispute of over the Senkaku/Diaoyu islands, the Ya'an earthquake, H7N9 bird flu, and Chinese president Xi Jinping's "Chinese Dream" proposal. On average, each trending topics lasts 4.6 days, with the shortest lasting 2 days and the longest 14 days. The average topic includes 19K tweets and 635K comments, with min/max tweets being 394/108K, and min/max comments being 538/3.1M.

These numbers reveal that only a small fraction of the 36.5M tweets in our dataset belong to trending topics. This results is not surprising: just like on Facebook and Twitter, the vast majority of content on Weibo is random chatter.

For our study, the most interesting trending topics are ones that are censored. Table 1 lists the details of the top 5 most *censored topics* during our measurement period, ranked by the percentage of tweets in the topic that were censored. We observe that these censored topics touch on many political issues, and that the magnitude of censorship is highly variable (ranging between 82% and 6%). Of the remaining 32 trending topics, 27 exhibit <2% censorship. Table 2 lists some of the keywords in the top 5 most censored topics, as well as the corresponding morphs of those words.

Some readers may be surprised by how few topics are censored on Weibo. One reason for the lack of censorship is that the Chi-

	Avg. Comments per Tweet	Avg. Comments per User	Total Comments	Unique Commentors	Unique Tweeters
ρ	0.012	0.033	0.077	0.036	0.198
p-value	0.874	0.677	0.323	0.644	0.011

Table 3: Spearman's ρ correlation between percentage of censored tweets vs. 5 topic-level variables.

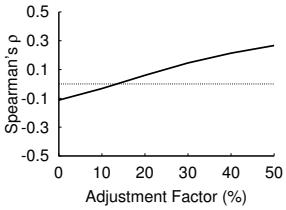

Figure 13: The Spearman's ρ between % of censored tweets vs. tweets per user.

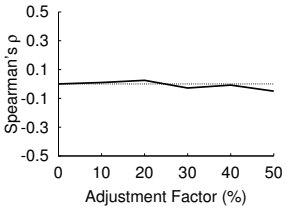

Figure 14: The Spearman's ρ between % of censored tweets vs. tweets per topic.

nese government is primarily interested in censoring content that incites public protests, not content that is critical of the government [25]. Thus, only a subset of the political discourse on Weibo is censored. Furthermore, political and news-related topics are a small percentage of the overall trending topics on social media. For example, only 17% of trending topics on Twitter are political or news-related [28]. Thus, people's limited appetite for political discussion on social media puts an upper bound on the number of topics that could potentially trigger censorship.

Impact of Censorship. We now examine the correlation between the magnitude of censorship and the characteristics exhibited by trending topics. We use Spearman's ρ for this analysis, which is a non-parametric measure of correlation between two variables. ρ is defined between -1 and 1, with $\rho > 0$ indicating positive correlation, and $\rho < 0$ indicating negative correlation. Prior work has successfully leveraged ρ to analyze correlations on social network datasets [12, 24].

Table 3 lists the correlation between the percentage of censored tweets in our 37 topics versus 5 other variables. To increase the sample size of our dataset, we divide each topic into separate days, *i.e.,* a topic with a lifetime of d days creates d separate daily samples. Thus, each test includes $n = 169$ samples. In each test, the null hypothesis is that the given variable is not impacted by censorship.

Table 3 shows that there is a weak positive correlation between the number of unique users who tweet in a topic and censorship. This suggests that Weibo users are not dissuaded from discussing sensitive topics by the threat of censorship. The other four variables presented in Table 3 do not show any correlation with censorship.

Next, we examine the correlation between censorship and the number of tweets per user. Analyzing tweets per user is challenging because our crawler misses some fraction of censored tweets (see § 3.4). To compensate, we *adjust* the number of tweets per user by assuming that some percentage of censored tweets were missed. For example, if we assume that 50% of censored tweets were missed, and user u generates 10 tweets in topic A, 2 of which are censored, then we estimate u actually generated 12 tweets in A.

Figure 13 shows the correlation between censorship and average tweets per user in our 37 topics. The x-axis denotes the *adjustment factor*, defined as the estimated percentage of censored tweets that were missed by the crawler. 0% adjustment refers to the original, unmodified average tweets per user. Interestingly, the unadjusted data shows negative correlation between censorship and tweets per user, suggesting that censorship may be dissuading users from tweeting. However, when the missing censored tweets are taken into account, the correlation quickly becomes strongly positive. This reveals that users actually generate more tweets than normal in censored topics.

Lastly, we examine the correlation between censorship and total tweets per topic. Tweets per topic is also impacted by missing censored tweets, so we apply the same adjustment methodology used in the previous experiment. Figure 14 shows that there is no clear correlation between censorship and total tweets per topic, even when the *adjustment factor* is taken into consideration.

Discussion. Our analysis reveals surprising aspects about the impact of censorship on Weibo. Initially, we assumed that censorship caused a chilling effect that would manifest as negative correlations, *i.e.,* fewer active users, fewer tweets per user, *etc.* Instead, the data reveals the opposite effects, *i.e.,* censored topics see more active users tweeting more frequently. As shown in Table 3, censorship does not correlate with reduced overall discussion volume, nor does it impact commenting behavior (probably because comments are not censored). These results indicate that, at least for our target population on Weibo during our measurement period, censorship does not cause a chilling effect on discussions.

Our results are different from those of Fu *et al.*, who found that the Real Name Registration (RnR) policy had a chilling effect on Weibo users [42]. However, RnR was implemented on March 16, 2012, and Fu *et al.* observe that the volume of tweets returned to normal levels by June 2012. Thus, it is possible that the chilling effect of RnR has dulled over time.

6. WORD USAGE ON WEIBO

In § 5, we quantify the impact of censorship on the high-level dynamics of trending topics on Weibo. This brings us to the second major question posed in this paper: *do Weibo users adapt in order to avoid censorship*?

To answer this question, we analyze the relationship between censorship and morph usage. As mentioned in § 5, 11 of our 37 trending topics include morphs, and prior work has also observed censored topics that include morphs [7, 42].We observe that the vast majority of morph usage in our trending topics occurs within heavily censored topics, indicating that there is a relationship between censorship and morph usage. To better understand this phenomenon, we examine the temporal usage of morphs, and the usage of morphs by different types of users (*e.g., celebrities*), in both censored and uncensored topics.

6.1 What is a Morph?

Before we analyze morphs, we must first clearly define what a morph is. A morph is an alternate form of a preexisting, original word or phrase. In conversation (online or offline), the morph can be substituted for the original.

We observe that some morphs on Weibo are existing words used as metaphors or for satirical effect. For example, some Weibo users refer to Chinese president Xi Jingping (习近平) as the "crown prince" (储君). Similarly, after several scandals involving the Red Cross (红十字) in China, Weibo users began referring to it as the Black Cross (黑十字). Alternatively, some morphs are generaliza-

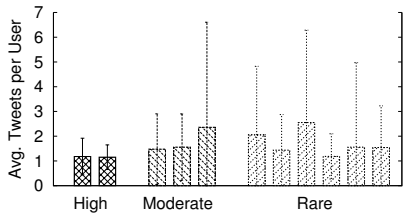

Figure 15: Average tweets per user for different topics.

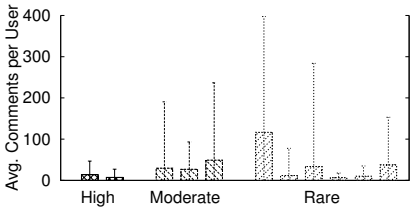

Figure 16: Average comments per user for different topics.

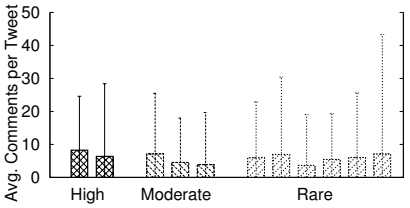

Figure 17: Average comments per tweet for different topics.

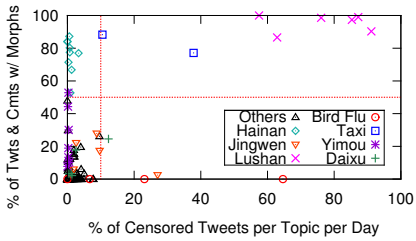

Figure 18: Morph usage versus censorship across 37 trending topics.

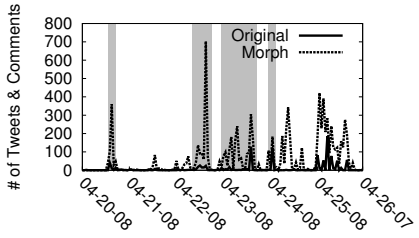

Figure 19: Original word and morph usage over time in the *Lushan* topic.

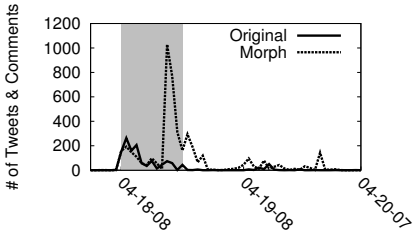

Figure 20: Original word and morph usage over time in the *Taxi* topic.

tions that only make sense in context. For example, in the *Taxi* topic in Table 2, users replaced the name of the taxi driver (郭立新) with the generic phrase "Beijing taxi driver" (北京的哥).

We also observe morphs that are entirely new words, or words used in unexpected ways. In the *Jingwen* topic in Table 2, users partially anglicanize 京温 to 京*wen*. In the same topic, some users replace the girl's name (袁利亚) with a homograph (袁莉亚), *i.e.,* a word that looks similar to the original. In a topic discussing the Ling Zhu Thallium poisoning case, some users replace the girl's name (朱令) with a (in this case, offensive) homophone (猪令), *i.e.,* a word that looks different but has the same sound as the original.

Identifying Morphs. In this section, our goal is to investigate whether users adapt to censorship by inventing new morphs. Thus, we are only interested in novel morphs that were invented within our 37 trending topics. To identify novel morphs, we had two native Chinese speakers identify all morphs in the lists of 100 words associated with each of our 37 topics. We then counted the number of tweets and comments using each morph during each day of our dataset. We assume that any a morph m from topic A used >100 times prior to the start date of A was not invented during A, and is therefor not a novel morph.

In total, we identified 11 trending topics in our dataset that include novel morphs. Table 2 lists the original words and morphs in the top 5 most censored topics. Note that a single original word can correspond with multiple morphs.

General Statistics. We now briefly present some general statistics about the 11 topics that include novel morphs. We divide the 11 topics into three categories: *high*, *moderate*, and *rare* censorship. The high-censorship category includes the *Lushan* and *Taxi* topics, the moderate category includes *Jingwen*, *Obama*, and *Zhuling*, and the rare category includes the remaining six topics.

Figure 15 through 17 show that all 11 topics have similar average tweets per user, comments per user, and comments per tweet. In some cases the standard deviation is quite high, particularly for comments in Figure 16. This is due to the presence of spam accounts within the topic. Note that the values in Figures 15 and 17 are not adjusted to compensate for censored tweets missed by the crawler (see § 3.4).

6.2 Morph Usage and Censorship

The first question we address in this section is: *is there a relationship between censorship and morph usage on Weibo*? To answer this question, we plot the percentage of tweets and comments that use morphs in each of our 37 topics, versus the percentage of tweets that were censored in each topic. Figure 18 shows the results of this experiment, where each point represents one day of tweets and comments from one topic. Topics above the 50% horizontal line use more morphs than original words on that particular day.

The majority of trending topics in our dataset do not include novel morphs and are not censored, thus they cluster at the origin in Figure 18. However, several topics exhibit different behavior. The *Hainan* topic, and for one day the *Yimou* topic, appear in the top left of Figure 18, *i.e.,* morphs dominate but the topics are uncensored. In both cases, these morphs are comedic in nature. In *Hainan*, users invented a pejorative term for certain kinds of women (绿茶婊). In *Yimou*, users equivocate the one child policy with a popular Chinese children's animation (葫芦娃).

Three trending topics appear in the bottom right quadrant of Figure 18, *i.e.,* they are censored but original words dominate over morphs. The first topic corresponds to the *Jingwen* suicide/murder case. The censored point in the bottom right quadrant represents to the first day of this topic. As shown in Table 1, users did invent morphs in the *Jingwen* topic. However these morphs were not used until the latter four days of topic, when it was not censored. The same trend of first-day censorship, followed by morph introduction happens in the *Daixu* topic (about a senior colonel who posted a controversial tweet about loss of life due to bird flu). The third topic also concerns *Bird Flu*. As shown in Table 1, users did not invent any morphs within this topic.

Finally, *Lushan* and *Taxi*, the two most heavily censored topics in our dataset appear in the upper right quadrant of Figure 18. The vast majority of tweets and comments in these two topics use morphs instead of the original words.

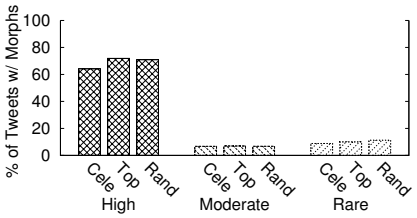

Figure 21: Morph usage in tweets by different user groups.

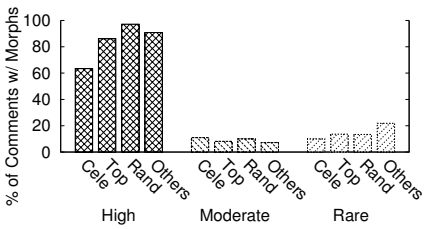

Figure 22: Morph usage in comments by different user groups.

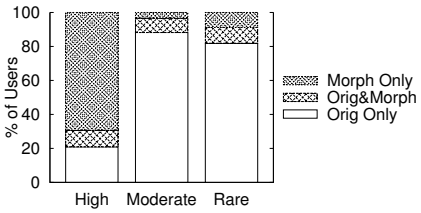

Figure 23: Use of original words and morphs by individual users.

Discussion. Figure 18 reveals that there is a direct relationship between censorship and morph usage. Morphs consistently dominate original words in *only one* of the 32 trending topics in our dataset with $\leq$5% censored tweets. In contrast, the two most censored topics are overwhelmingly dominated by morphs. Two of the three topics in the middle (between 5 and 20% censorship) include morphs, but usage is weighted towards original words.

One concern with Figure 18 is that our dataset is missing some censored tweets (see § 3.4). In heavily censored topics, it is possible that we may not observe many censored tweets that include original words. However, recall that there are an order of magnitude more comments than tweets on Weibo (§ 4.2), and comments are not censored. The results in Figure 18 primarily derive from word usage in comments. Thus, the results in Figure 18 are not significantly impacted by censored tweets that are missing from our dataset.

6.3 Dynamics of Morph Usage

Figure 18 indicates that users adopt morphs as a means to avoid censorship. In this section, we examine the dynamics of morph usage and adoption in greater detail.

Morph Usage Over Time. First, we examine how the use of morphs changes over time. Figures 19 and 20 plot the number of tweets and comments that use original words or morphs per hour in our top two most censored topics. Grey regions denote times when censorship was occurring. Although the *Lushan* topic is very bursty, it can be seen that the morphs exist at the start of the topic on April 20, 2013. The same observation is true for the *Taxi* topic, although the trend is clearer: initially, the original word and the morph are equally popular. However, 10 hours after tweets with the original words are censored, the popularity of the morph spikes, while the original words fall out of favor.

There are two takeaways from the results in Figures 19 and 20. First, the morphs are invented at essentially the same time the topics begin to trend, *i.e.,* users preemptively invent morphs, even before there are signs of censorship. This may indicate that Weibo users take a proactive approach to inventing morphs that can be used to avoid keyword-based censorship. Second, Figure 20 suggests that the popularity of morphs can skyrocket due to censorship of the original words. However, given that we have only observed one example of this phenomenon, we cannot rule out that some external factor is the cause of the popularity spike in Figure 20.

Morph Usage by Different Types of Users. Next, we examine how morph usage differs across different types of users. Figures 21 and 22 plot the percentage of tweets and comments that use morphs from *celebrities*, *top commentors*, and *random commentors*. As in § 6.1, the 11 topics are divided into three groups based on the censorship rate. In Figure 22, there is an additional bar for "other" users who were not crawled, but did comment on tweets we crawled.

Figures 21 and 22 reinforce our finding that morph usage is correlated with censorship. Across all user groups, morphs appear in 63-97% of tweets and comments in highly censored topics. Conversely, morph usage is very uncommon even in moderately censored topics.

The difference between Figures 21 and 22 is that morphs are more common in comments, especially among non-celebrities. We hypothesize that this occurs for two reasons. First, non-celebrities may be more willing to experiment with novel morphs whose meaning is not known to a large audience. Conversely, *celebrities* may be less willing to use novel morphs that may not be understood by their audience. Second, because tweets create a context for their attached comments, it may be easier to use novel morphs in comments. For example, if a tweet discusses Xi Jinping, it is obvious who the "crown prince" in comments refers to.

Evolution of Word Use by Individuals. Next, we seek to understand whether individual users adapt their word usage over the course of a trending topic. To analyze this, Figure 23 plots the percentage of users who use only original words, only morphs, or a combination of the two. We observe that the number of users who use both original words and morphs is only $\approx$9%, regardless of the magnitude of censorship of the topic. This shows that most users do not alter their word usage, *i.e.,* users choose the convention they will use when they first tweet/comment, and they do not deviate from this convention though the life of the topic.

Instead, the spikes in morph popularity observed in Figures 19 and 20 are due to *communal* adaptation. As more users join these trending topics over time, they chose to adopt either the original word or the morph convention. In the case of high-censorship topics, Weibo users joining the conversation overwhelmingly choose to adopt morphs, possibly because those tweets are less likely to be censored.

Word Correspondence. Finally, we examine the correspondence between the words used in tweets and their associated comments. We seek to answer the question: *do commentors adopt the conventions used in the associated tweet*? To answer this question, Figure 24 plots the percentage of comments that use the *same* con-

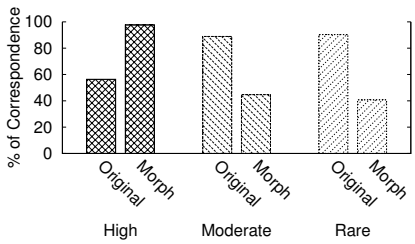

Figure 24: Correspondence between words used in tweets and their associated comments.

vention as their associated tweet, *i.e.,* comments that use the original word on tweets that also use the original word, and vice-versa.

Figure 24 reveals that the word correspondence trends are opposite for highly censored topics versus moderately and rarely censored topics. In highly censored topics, 44% of comments use morphs even when the tweet uses the original word. Conversely, in low-censorship topics, only 10% of comments use morphs when the tweet uses the original word. The trend reverses when we consider tweets that use morphs: in highly censored topics, commentors eagerly adopt the morph convention, whereas in low-censorship topics, commentors revert to using the original words.

The results in Figure 24 indicate that the conventions adopted by commentors are influenced by censorship. In high-censorship topics, commentors tend to use morphs regardless of the convention used by the tweet. Comments that use novel morphs on tweets that use original words help to establish the context of the morph for users who have not observed it before, which may help speed the adoption of the morph. In contrast, commentors are reluctant to adopt novel morphs in the absence of censorship. Without the impetus of censorship, commentors revert to using original words even when tweets include novel morphs.

7. RELATED WORK

Information Dissemination. Information dissemination on OSNs has been extensively studied in the literatures. Many studies focus on Twitter: [37] measure retweets, while [31, 40, 36, 13, 29] measure $\#hashtags$. [10, 12] investigate the impact of social influence on Facebook and Twitter. Numerous studies have applied machine learning algorithms to the prediction of trending topics and user attributes based on information dissemination patterns [33, 30, 46, 19].

Topic Models on OSNs. Several studies share the aim of extracting topics from OSN data. [34, 35] leverage Labeled LDA to extract topics from short tweets. [11] use a graph-based approach to identify emerging topics. Lastly, [20] evaluate the efficacy of several topic models (*e.g.,* TF*IDF) on Twitter data. In this work, we leverage LDA for topic extraction. Unlike Twitter, LDA is successful on Weibo because there are more words per tweet, and comments can be used to increase the length of each document.

Linguistic Evolution in Social Media. Several studies delve into how linguistic conventions change over time on social media. [14, 15, 16] study linguistic style accommodation, power differentials between users revealed through linguistics, and linguistic change over long time scales in social media. [27, 26] measure and predict the emergence of social conventions on Twitter. In our study, we also investigate the emergence of conventions (*i.e.,* morphs). However, as shown in § 6, morphs emerge over very short time scales in response to censorship, which is a different environment than that studied by prior work. [21] propose a graph-based approach for identifying the original word associated with novel morph on Weibo.

8. CONCLUSION

In this paper, we study the impact of censorship on discourse and word choice in Sina Weibo. We crawled 280K Weibo users on an hourly basis for 44 days, gathering 839M tweets and comments. Our study is the first to analyze comments on Weibo, which is crucial since there are an order of magnitude more comments than tweets. We observe that ≈1% of all tweets are censored, although some topics are 82% censored, and some celebrity users are 50% censored.

Our analysis of trending topics reveals that there are positive correlations between censorship and user engagement. This may indicate that censorship has less of a chilling effect on discourse on Weibo than was previously suspected [42]. However, we caution that estimating the magnitude of the chilling effect, or lack thereof, is difficult given our data: although we observe many users discussing sensitive topics, it is possible that even more users would discuss these topics if there was no threat of censorship. Thus, although we observe positive correlations between censorship and user engagement, these correlations could be even higher in the complete absence of censorship.

We also observe a strong relationship between censorship and the use of morphs. Weibo users tend to introduce novel morphs into heavily censored topics within the first few hours of the topics existence, even before censorship has been implemented. This indicates that users are aware of censorship and actively adopt novel morphs as a way to avoid keyword censorship.

Taking a broader view, we see this study as a first step towards understanding the impact of censorship on discourse in social media. There are several future directions that could strengthen and extend our findings. *First*, it would be beneficial to confirm our findings over longer time scales (this study only examines two months of data) and through more diverse socio-political conditions (*e.g.,* elections, leadership changes, and natural disasters).

Second, this study analyzes the impact of censorship on the macro-scale, aggregate behavior of Weibo users. Additional work is necessary in order to understand the micro-scale dissemination of morphs through the Weibo population. Unfortunately, we cannot conduct this analysis on our Weibo dataset due to the confounding impact of comments. For example, user A can disseminate a morph to user B's followers by commenting on B's tweet, even if there are no social links between A and those followers. New information dissemination models that take these indirect information channels into account will need to be developed before we can model the dissemination of morphs on Weibo.

Lastly, we note that the point of our study is not to make value judgments for or against censorship. Our goal is simply to observe the impact these policies have on social media, as a step towards improving models and predictors of information dissemination and linguistic change.

Acknowledgments

We thank the anonymous reviewers and our shepherd, Krishna Gummadi, for their valuable time and comments. This research was supported by an Amazon Web Services in Education Grant.

9. REFERENCES

[1] Sensitive words series. China Digital Times. http://chinadigitaltimes.net/china/sensitive-words-series/.
[2] 2013年4月份网络热点事件舆情报告 (Trending Topics Report in April 2013). Mesh Media, 2013.
[3] 2013年5月份网络热点事件舆情报告 (Trending Topics Report in May 2013). Mesh Media, 2013.
[4] Sogou pinyin. Sogou, 2013. http://pinyin.sogou.com/.
[5] Sohu.com inc sohu q1 2013 earnings call transcript. Morningstar, Inc, 2013.
[6] 新浪微博社区公约 (Sina Weibo Community Treaty). Sina, 2013.
[7] BAMMAN, D., O'CONNOR, B., AND SMITH, N. A. Censorship and Deletion Practices in Chinese Social Media. *First Monday 17*, 3 (2012).
[8] BENEVENUTO, F., MAGNO, G., RODRIGUES, T., AND ALMEIDA, V. Detecting spammers on twitter. In *Proc. of CEAS* (2010).

[9] BLEI, D. M., NG, A. Y., AND JORDAN, M. I. Latent dirichlet allocation. *the Journal of machine Learning research 3* (2003).

[10] BOND, R. M., FARISS, C. J., JONES, J. J., KRAMER, A. D. I., MARLOW, C., SETTLE, J. E., AND FOWLER, J. H. A 61-million-person experiment in social influence and political mobilization. *Nature 2012* (489).

[11] CATALDI, M., DI CARO, L., AND SCHIFANELLA, C. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proc. of MDMKDD* (2010).

[12] CHA, M., HADDADI, H., BENEVENUTO, F., AND GUMMADI, K. P. Measuring user influence in twitter: The million follower fallacy. In *Proc. of ICWSM* (2010).

[13] CUNHA, E., MAGNO, G., COMARELA, G., ALMEIDA, V., GONÇALVES, M. A., AND BENEVENUTO, F. Analyzing the dynamic evolution of hashtags on twitter: a language-based approach. In *Proc. of LSM* (2011).

[14] DANESCU-NICULESCU-MIZIL, C., GAMON, M., AND DUMAIS, S. Mark my words! Linguistic style accommodation in social media. In *Proc. of WWW* (2011).

[15] DANESCU-NICULESCU-MIZIL, C., LEE, L., PANG, B., AND KLEINBERG, J. M. Echoes of power: Language effects and power differences in social interaction. In *Proc. of WWW* (2011).

[16] DANESCU-NICULESCU-MIZIL, C., WEST, R., JURAFSKY, D., LESKOVEC, J., AND POTTS, C. No country for old members: User lifecycle and linguistic change in online communities. In *Proc. of WWW* (2013).

[17] GAO, H. Rumor, lies, and weibo: How social media is changing the nature of truth in china. The Atlantic, 2012.

[18] GRIER, C., THOMAS, K., PAXSON, V., AND ZHANG, M. @spam: the underground on 140 characters or less. In *Proc. of CCS* (2010).

[19] GUO, L., TAN, E., CHEN, S., ZHANG, X., AND ZHAO, Y. E. Analyzing patterns of user content generation in online social networks. In *Proc. of KDD* (2009).

[20] HONG, L., AND DAVISON, B. D. Empirical study of topic modeling in twitter. In *Proc. of SOMA* (2010).

[21] HUANG, H., WEN, Z., YU, D., JI, H., SUN, Y., HAN, J., AND LI, H. Resolving entity morphs in censored data. In *Proc. of ACL* (2013).

[22] INC., R. Renren announces unaudited third quarter 2012 financial results. PRNewsWire, 2012.

[23] JADESOUL. Open Chinese Lexical Analysis System. Github, 2013. `https://github.com/jadesoul/openclas`.

[24] JIANG, J., WILSON, C., WANG, X., HUANG, P., SHA, W., DAI, Y., AND ZHAO, B. Y. Understanding latent interactions in online social networks. In *Proc. of IMC* (2010).

[25] KING, G., PAN, J., AND ROBERTS, M. E. How censorship in china allows government criticism but silences collective expression. *American Political Science Review* (2013).

[26] KOOTI, F., MASON, W. A., GUMMADI, P. K., AND CHA, M. Predicting emerging social conventions in online social networks. In *Proc. of CIKM* (2012).

[27] KOOTI, F., YANG, H., CHA, M., GUMMADI, P. K., AND MASON, W. A. The emergence of conventions in online social networks. In *Proc. of ICWSM* (2012).

[28] LEE, K., PALSETIA, D., NARAYANAN, R., PATWARY, M. M. A., AGRAWAL, A., AND CHOUDHARY, A. Twitter trending topic classification. In *Proc. of IEEE ICDMW* (2011).

[29] LEHMANN, J., GONÇALVES, B., RAMASCO, J. J., AND CATTUTO, C. Dynamical classes of collective attention in twitter. *CoRR* (2011).

[30] LESKOVEC, J., BACKSTROM, L., AND KLEINBERG, J. Meme-tracking and the dynamics of the news cycle. In *Proc. of KDD* (2009).

[31] MA, Z., SUN, A., AND CONG, G. Will this #hashtag be popular tomorrow? In *Proc. of SIGIR* (2012).

[32] MARTIN, R. China internet penetration map 2012. Tech In Asia, January 2012. `http://www.techinasia.com/china-internet-penetration-map/`.

[33] PENNACCHIOTTI, M., AND POPESCU, A.-M. Democrats, republicans and starbucks afficionados: user classification in twitter. In *Proc. of KDD* (2011).

[34] QUERCIA, D., ASKHAM, H., AND CROWCROFT, J. Tweetlda: Supervised topic classification and link prediction on twitter. In *Proc. of WebSci* (2012).

[35] RAMAGE, D., DUMAIS, S. T., AND LIEBLING, D. J. Characterizing microblogs with topic models. In *Proc. of ICWSM* (2010).

[36] ROMERO, D. M., MEEDER, B., AND KLEINBERG, J. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proc. of WWW* (2011).

[37] STARBIRD, K., AND PALEN, L. (how) will the revolution be retweeted?: information diffusion and the 2011 egyptian uprising. In *Proc. of CSCW* (2012).

[38] STRINGHINI, G., KRUEGEL, C., AND VIGNA, G. Detecting spammers on social networks. In *Proc. of ACSAC* (2010).

[39] THOMAS, K., ET AL. Suspended accounts in retrospect: An analysis of twitter spam. In *Proc. of IMC* (2011).

[40] TSUR, O., AND RAPPOPORT, A. What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In *Proc. of WSDM* (2012).

[41] WA FU, F., AND CHAU, M. Reality Check for the Chinese Microblog Space: A Random Sampling Approach. *PLoS ONE 8*, 3 (2013).

[42] WA FU, K., HONG CHAN, C., AND CHAU, M. Assessing censorship on microblogs in china: Discriminatory keyword analysis and the real-name registration policy. *IEEE Internet Computing 17*, 3 (2013).

[43] WALLACH, H. M. Structured topic models for language. *Unpublished doctoral dissertation, Univ. of Cambridge* (2008).

[44] WENLIN, Z. Weibo has over 300 million users, and 100 million tweets daily. Xinhua News, February 2012.

[45] WENLIN, Z. Weibo has over 500 million users, and 4.6 million active users daily. Xinhua News, February 2013.

[46] ZAMAN, T. R., HERBRICH, R., GAEL, J. V., AND STERN, D. Predicting information spreading in twitter. In *Proc. of NIPS* (2010).

[47] ZHANG, H.-P., YU, H.-K., XIONG, D.-Y., AND LIU, Q. HHMM-based Chinese lexical analyzer ICTCLAS. In *Proc. of SIGHAN workshop on Chinese language processing* (2003).

[48] ZHOU, M. 微博明日起实行实名制 用户凭有效信息才可注册 (Weibo RnR Launching Tomorrow: Users Need to Register with Valid Information). 凤凰科技讯, March 2013.

[49] ZHU, T., PHIPPS, D., PRIDGEN, A., CRANDALL, J. R., AND WALLACH, D. S. The velocity of censorship: High-fidelity detection of microblog post deletions. In *Proc. of USENIX Security* (2013).