Comparing the Pulses of Categorical Hot Events in Twitter and Weibo

Xin Shuai
Dept. of Informatics
Indiana University
Bloomington
xshuai@indiana.edu

Xiaozhong Liu
Dept. of Information and Library Science
Indiana University
Bloomington
liu237@indiana.edu

Tian Xia
Dept. of Information Resource Management
Renmin University of China
xiatian1119@gmail.com

Yuqing Wu
Dept. of Computer Science
Indiana University
Bloomington
yuqwu@indiana.edu

Chun Guo
Dept. of Information and Library Science
Indiana University
Bloomington
chunguo@indiana.edu

ABSTRACT

The fragility and interconnectivity of the planet argue compellingly for a greater understanding of how different communities make sense of their world. One of such critical demands relies on comparing the Chinese and the rest of the world (e.g., Americans), where communities' ideological and cultural backgrounds can be significantly different. While traditional studies aim to learn the similarities and differences between these communities via high-cost user studies, in this paper we propose a much more efficient method to compare different communities by utilizing social media. Specifically, Weibo and Twitter, the two largest microblogging systems, are employed to represent the target communities, i.e. China and the Western world (mainly United States), respectively. Meanwhile, through the analysis of the Wikipedia page-click log, we identify a set of categorical 'hot events' for one month in 2012 and search those hot events in Weibo and Twitter corpora along with timestamps via information retrieval methods. We further quantitatively and qualitatively compare users' responses to those events in Twitter and Weibo in terms of three aspects: popularity, temporal dynamic, and information diffusion. The comparative results show that although the popularity ranking of those events are very similar, the patterns of temporal dynamics and information diffusion can be quite different.

Categories and Subject Descriptors

J.4 [**Social and Behavioral Sciences**]: Sociology; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Measurement

HT'14, September 1–4, 2014, Santiago, Chile.

http://dx.doi.org/10.1145/2631775.2631810.

Keywords

Social Media; Information Retrieval; Community Comparison; Twitter; Weibo; Wikipedia; Click Log Mining; Information Diffusion

1. INTRODUCTION

The way people perceive and exploit their cultural environments through the social media has been observed and well documented [6, 13, 21], including sentiment analysis [1, 15], the potential role of the Internet in China [3], comparisons of decision-making in face-to-face versus computer-mediated communication, network influences on social isolation [22], predictions about the role of media on society [21], as well as instrumental uses of Twitter as a communication tool [7]. However, to the best of our knowledge, few scholarly studies ever conducted comprehensive comparisons of users' behavior in social media in China and the Western world (mainly United States), especially during the periods of hot events. For social scientists, such comparisons are becoming increasingly crucial and intriguing, because the responses of users from the two types of social media are quite representative of Chinese and Americans, respectively. With a new, compelling global landscape being cultivated world-wide [24] due to the growth of China in economic, political, and cultural aspects, mining and comparing large scale datasets from China and United States can help us better understand the ideological and cultural differences between the two of the world's powers.

Social media, especially microblogging systems, can efficiently and effectively reflect real world events [1], which provide good dynamic recourses for researchers to conduct various studies of the large scale of users or communities at a low cost, including information diffusion [17], information ranking [11, 12], sentiment analysis [15], and social networks analysis [8]. Motivated by these findings, in order to compare two of the largest microblogging user groups in the world, we collect massive Twitter and Weibo corpora for comparative studies with a number of innovative indicators. Although Twitter and Weibo are both microblogging platforms with very similar functionalities, they are consumed by totally different users: One, the default languages of Twitter and Weibo are English and Chinese, respectively; Two, Twitter access is strictly forbidden in mainland China due to political reasons [23], while, theoretically, the whole world can access to the Weibo platform. However,

as Weibo's default language is Chinese, the primary users are people of Chinese heritage, even though they may physically reside anywhere in the world.

Although accessing Twitter is an impossible mission in China, an increasing number of Chinese users began to seek the real-time information from the rest of the word via other channels, like news, search engines, and other websites. With a very different cultural and ideological background, Chinese users' reaction and interest toward the same event or topic could be different (or similar) from the rest of the world (e.g., US). Fortunately, Twitter and Weibo provide us a good opportunity to investigate and compare the two groups of users by analyzing their textual messages published on the two microblogging platforms.

The main goal of this paper is to compare the similarities and differences in response to hot events between Twitter and Weibo users from textual and social network perspectives. The contributions of this paper are twofold:

- One, we identify a set of "Hot Events" via peak detection and trend analysis from Wikipedia click log. In addition, based on Wikipedia category metadata, we group all the hot events into different categories, such as Science, Politics, and Sports, etc. We, then, trace the temporal pulses of categorical hot events in both Twitter and Weibo corpora utilizing information retrieval methods. Especially, Wikipedia offers both Chinese and English content and metadata for each candidate event, which enables cross-language search and mining for both Twitter and Weibo corpora.

- Two, from textual and social network perspectives, we propose several indicators to compare Weibo and Twitter response towards the same set of categorical hot events. We apply statistical analysis and case study methods to both quantitatively and qualitatively compare the two communities.

Experiment result shows that, while Twitter and Weibo communities share similar interests from the event popularity perspective, temporal analysis and information diffusion modeling reveal that Twitter and Weibo users are different in consuming those hot events. Especially, we find Weibo and Twitter users are more similar when they are contagious to the hot events in Science and Technology categories. On the contrary, some other categories, i.e., Arts and Politics, distinguish Twitter from Weibo users.

In the remainder of this paper, Section 2 reviews relevant literature and methodology for social media mining and comparison, Section 3 proposes our novel method for comparing Twitter and Weibo communities, Section 4 describes the experiment setting and evaluation results, and Section 5 discusses the findings and limitations of the study and identifies subsequent research steps.

2. LITERATURE REVIEW

Information retrieval and text mining algorithms are used by scholars to analyze and compare large textual corpora, especially to investigate users' interest [11] via sentiment analysis [15]. In this context, Twitter and Weibo, the most popular microblogging systems, have been successfully used to represent and investigate Western (mainly US) and Chinese communities, respectively. For instance, Baucom et al. [1] used Twitter to mirror real world events and found that Twitter sentiment along with geo-location information can be used to estimate very dynamic real world events, e.g., score changes in athletic events. Similar studies [2,15,18] also verify the close relatedness between real-world events and chronological Twitter data. It is clear that massive Twitter data can be used to characterize and predict the real-world events, which has been successfully applied to a number of data mining tasks, i.e., information retrieval [19], information diffusion [17], and event prediction [16].

In the past few years, the number of Chinese Internet users is growing very fast. So far, more than 20% of Internet users come from China, thus investigating the behavior of Chinese Internet users becomes increasingly important [3, 24]. While using Twitter to characterize real world events is well documented, Weibo is becoming an important means to understand the Chinese community. For instance, Zhao et al., [25] employed Weibo data to investigate event discussion by using term-message-user network. They used random-walk algorithms to study the temporal event information diffusion, and the event is pre-defined by domain expert. Similarly, Guan et al., [5] studied 21 (expert pre-defined) hot events of Weibo by utilizing 32 prestigious users (influential users).

Unfortunately, due to the language and political barriers, most users from each community can only access one system exclusively. While most previous studies treat Twitter and Weibo as the same kind of social media except for language, some other researchers [4, 10] found that Chinese Weibo may have some unique features. Not until recently, some researchers became aware of the importance of comparing the Weibo and Twitter corpora. For instance, Gao et al., [4] compared Twitter and Weibo corpora from sentiment, entity, system access perspectives. A list of comparison indicators were listed in the Table 1.

Table 1: Twitter and Weibo comparison in previous studies

Comparison Indicator	Previous Studies	Findings
HashTag distribution	[4, 10]	Weibo users are interested in entertainment and sports topics, and Weibo users like more joke related content comparing with Twitter users.
URL distribution	[4]	Weibo users post less URLs compared with Twitter users.
Forward distribution	[10]	Weibo users forward message slower than Twitter.
Follow distribution	[10]	Twitter users number of actions will have a more significant effectiveness on the number of "Followers" than that of Weibo.
Gender distribution (for 32 users)	[5]	Male users are more likely to be involved.
Picture distribution (for 32 users)	[5]	Messages containing pictures are more likely to be posted.
Sentiment distribution	[4]	Weibo users post more positive messages comparing with Twitter users.
System access distribution	[4]	On Twitter, more than 95% of the users use more than one client application while on Sina Weibo around 65% of the users switch between different clients.
Entity distribution	[4]	Weibo users post more entity information than Twitter users.

All of these comparative studies are inherently similar; they all focus on comparing some statistical properties of microblogging features, like HashTag, forwarding linkage, following linkage, etc. While those are all very interesting findings, they provide very limited knowledge about the differences and similarities between China and United States in the real world. To be specific, no prior study ever investigated topical or categorical Twitter and Weibo comparison during hot events, which is important; the nature of Weibo (or Twitter) users' responses to e.g. *Political* news can be very different from that of *Science*.

Different from these studies, this paper paves a new way to investigate the similarity and difference between Weibo and Twitter at the topical level, exploring the categorical hot events extracted from the Wikipedia page click log. We propose multiple indicators

to compare Twitter and Weibo communities. Meanwhile, information diffusion techniques [14, 17] are used in this study for social network-based comparison.

3. RESEARCH METHODS

The overall framework of our study is shown in Figure 1 and can be decomposed into four steps: First, the pre-processing of textual messages in Twitter and Weibo in different languages; Second, the identification of categorical hot events from the Wikipedia dump and click logs; Third, the tracing of hot event pulses in Twitter and Weibo corpora via information retrieval; Fourth, the comparison of Twitter and Weibo hot event pulses in terms of event category and a number of other indicators.

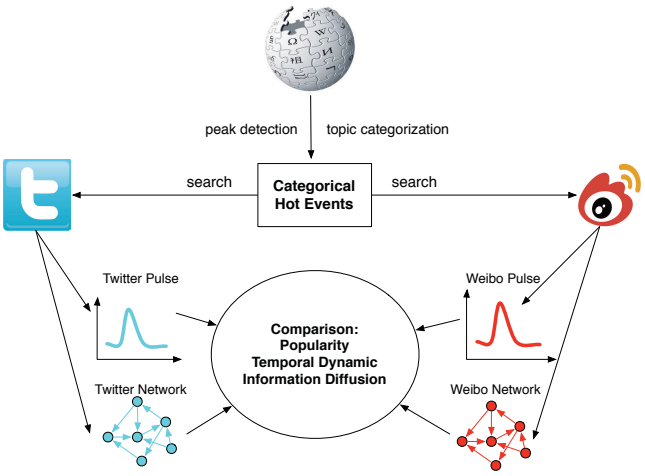


Figure 1: Twitter and Weibo comparison via Wikipedia categorical hot events.

3.1 Twitter and Weibo pre-processing

To pre-process the textual messages from Twitter and Weibo is to index all those messages to support full-text search. Specifically, we index the following fields for each message: id, creator, timestamp, content, and hashtags. The most important step in indexing is word tokenization. Since the English and Chinese languages are totally different, we apply different tokenization techniques to Twitter and Weibo, respectively.

For Twitter, we split the sentence into word tokens at white spaces and punctuation symbols, remove all stop words according to [9], and convert all tokens into lowercase. In addition, we apply stemming techniques to normalize the form of tokens.

For Weibo, since the Chinese language contains both simplified and traditional versions and does not delimit words by white spaces, tokenization is more challenging than in Twitter. We first normalize the words by converting all traditional characters into simplified characters, then apply the CRF model [20] to segment Chinese sentences into tokens.

3.2 Hot Event Identification via Wikipedia Click Log

In order to compare the responses of Twitter and Weibo communities to hot events, we first identify a set of candidate hot events. In this study, we discover hot events from Wikipedia. As the largest online encyclopedia, Wikipedia has become the most common online resources to gain knowledge and information about the world for people around the globe. When some hot events occur, people tend to view the Wikipedia pages relevant to those events, thus generating a traffic spike in the click logs of those pages. Therefore, we can utilize Wikipedia as a proxy to sense what happens in the real world, and estimate the start and end time of hot events based on when traffic spikes occur. In particular, Wikipedia provides multi-language versions (e.g., English and Chinese) of the same page to facilitate the access of users from different countries. In addition, each Wikipedia page contains category metadata defined by page editors, so we're able to categorize events by analyzing the Wikipedia page category.

3.2.1 *Peak Detection in Wikipedia Click Log*

Wikipedia page view statistics provide the number of times a particular page has been viewed (i.e. clicked). Wikipedia provides hourly page view statistics about how many times each Wikipedia page has been clicked for each hour. We can easily aggregate the hourly click statistics to obtain daily click statistics for each page.

Some Wikipedia pages related to some real world events are likely to be viewed much more frequently during the time when those events receive media coverage than the time when they are little discussed by the public. This sometimes explains spikes in the click log statistics during periods of time when such events are taking place and receiving public attention. For example, those pages regarding to athletes who attend Olympics will be clicked more frequently during the Olympics than other times; Articles on topics pertaining to a particular holiday may get more hits around the time of year when the holiday takes place; During an election year, anything somehow related to that election may be viewed more than at other times.

Given a Wikipedia page p and a date range $[d_0, d_n]$, we use $T(p, d)$ to denote the number of clicks Wikipedia page p received on date d. To detect a peak in $\{T(p, d)|d \in [d_0, d_n]\}$, We define a threshold as:

$$h_p(d_0, d_n) = \underset{d\in[d_0,d_n]}{mean} T(p, d) + \alpha * \underset{d\in[d_0,d_n]}{std} T(p, d) \quad (1)$$

where $\underset{d\in[d_0,d_n]}{mean} T(p, d)$ and $\underset{d\in[d_0,d_n]}{std} T(p, d)$ is the mean and standard deviation of $T(p, d)$ over $[d_0, d_n]$. The threshold α determines the degree of "peakiness" sufficient to detect an event. We will manually set it based on some some preliminary observations and the tuning of α is left for future work

If we can find $d_p \in [d_0, d_n]$ such that $T(p, d_p) > h_p(d_0, d_n)$, we claim that there exists a spike $T(p, d_p)$ in the click logs of page p occurring on d_p, and some event relevant to page p was occurring during d_0 and d_n. If multiple values of d_p satisfy Equation 1, we choose the d_p such that $T(p, d_p) > T(p, d)$ for any $d \in [d_0, d_n]$ to be the peak.

Once p_d of the hot event is detected, we further determine the start and end date of the event. Specifically, we define the start date d_s and end date d_e as the latest date that satisfies $T(p, d_s) \leq \underset{d\in[d_0,d_n]}{mean} T(p, d)$, and the earliest date that satisfies $T(p, d_e) \leq \underset{d\in[d_0,d_n]}{mean} T(p, d)$, respectively. To fully capture the whole period of the event, we empirically apply three days offset before d_s, and after d_e. Finally, we obtain an event $E_p(d_s, d_p, d_e)$ related to Wikipedia page p, starting from d_s and ending at d_e with a peak on d_p. Note that the optimal selection of the number of offset days is left for future work.

3.2.2 *Event Categorization*

After identifying all $E_p(d_p, d_s, d_e)$, we categorize them into different topics. Since each event corresponds to a Wikipedia page, we can easily categorize those events by categorizing those pages. The

category system of Wikipedia is organized as an overlapping 'tree', formed by linking related categories in a hierarchy. Any category may branch into subcategories, and it is possible for a category to be a subcategory of more than one 'parent' category (A is said to be a parent category of B when B is a subcategory of A).

At the bottom of each Wikipedia page, a set of category names are provided. The page editor either uses existing category names, or creates a new category name and links that category to one of the existing category names. Generally, those user-defined categories are either too vague or too specific to serve as feasible options for the topical classifications as we need. Fortunately, Wikipedia has already identified 25 categories as the main topic classifications for all Wikipedia pages.[1] The overall category structure in Wikipedia is shown in Figure 2.

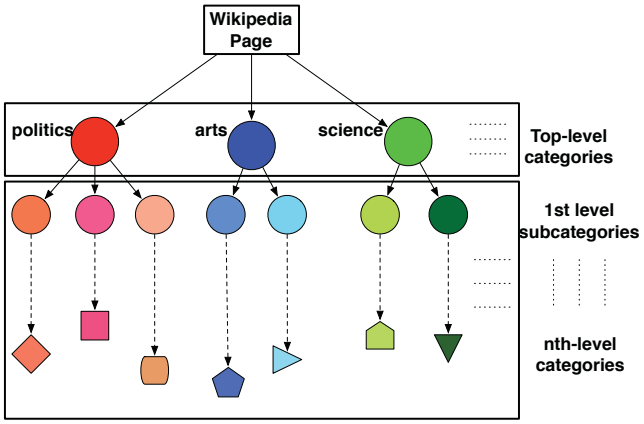


Figure 2: Wikipedia category hierarchy.

To categorize all Wikipedia page into one of the 25 categories, we create a vector with 25 dimensions for each page, with every dimension corresponding to one category. Then we take two steps to fill the values of the vector per dimension for each page. First, for each user-defined category we find at the bottom of the page, we trace it back in the tree-like category structure (see Figure 2) to find its corresponding top-level category; Second, we calculate TF-IDF score for each dimension of the vector. Finally, we pick the corresponding category with the highest TF-IDF score, as the most representative category for that page.

Table 2 outlines nine most important and representative categories (with event count) from the 25 categories. We conduct comparison analysis and illustrate experimental results based on these nine selected categories.

Table 2: Event category descriptions

Category	Exemplar Wikipedia pages	Counts
Agriculture(AGR)	Spinach, Cranberry, Royal jelly	63
Arts(ART)	Russell Crowe, Breaking Dawn, Yo-Yo Ma	114
Business(BUS)	Citigroup, McAfee, Loan	45
Culture(CUL)	Nobel Prize, Divine Comedy, Kung fu	93
Geography(GEO)	Times Square, Hawaii, Taipei	60
Politics(POL)	Iran, Barack Obama, Communist Party	181
Science(SCI)	Marie Curie, Atkins diet, Tsunami	95
Sports(SPO)	Table tennis, Miami Heat, FIFA World Cup	40
Technology(TEC)	Laser, 5G, Xbox	42

[1]http://en.wikipedia.org/wiki/Category:Main_topic_classifications

3.3 Tracing Hot Event via Information Retrieval

After locating categorical hot events from the Wikipedia corpus, we trace the hot event pulses from both the Twitter and Weibo corpora. Note that each Wikipedia page (we selected) has both Chinese and English titles provided by Wikipedia editors. For this study, we apply information retrieval techniques to search the Wikipedia title (English or Chinese) in the Twitter and Weibo indexes we built. In order to find accurate and comprehensive relevant messages from both indexes, we use title content exact match (some very short and ambiguous titles are removed). Unlike other information retrieval studies, ranking in this study is not important, since we care more about the daily total number of relevant messages, which indicates the hot event pulse (event mention probability) in Twitter or Weibo. However, there is one limitation to this approach: Wikipedia and Twitter or Weibo are written in different styles. For instance, both Weibo and Twitter limit the total length of messages, and users are more likely to use abbreviations or spoken language to represent the target event or topic. Therefore, only using Wikipedia titles to query relevant messages from Twitter and Weibo will result in low recall value.

To cope with this problem, for each Wikipedia page we expand the search query by incorporating all relevant redirect pages (in Chinese and English). For instance, for the Wikipedia page "Patient Protection and Affordable Care Act" we also use the titles of its redirected Wikipedia pages as the search queries, i.e., "Obamacare". Most of the time, users will write "obamacare" in the Twitter message, instead of "Patient Protection and Affordable Care Act". So, for each hot event, we send the following boolean query:

$$(\text{“}Page\ Title\text{”}\ OR\ \text{“}Redirect\ Page\ Titles\text{”})\ AND\ (Date\ ranging\ from\ Event\ Start\ Date\ to\ Event\ End\ Date) \quad (2)$$

The query indicates that all the relevant messages, in either Twitter or Weibo, should match at least one Wikipedia page title in the collection of event Wikipedia pages or any related redirected Wikipedia pages. Meanwhile, all the relevant messages should fall between event start and end dates.

After querying the Wikipedia page title plus redirected titles relevant to event $E_p(d_s, d_p, d_e)$, we come up with three types of temporal pulses: Wikipedia pulse, Twitter pulse and Weibo pulse. Specifically, given the same event, Wikipedia pulse represents the general public's attitude, Weibo pulse represents the Chinese attitude, while Twitter represents the American attitude (although people from most countries can get access to Twitter, the dominant users are still Americans). We use three time-series vectors with the same length to denote three types of temporal pulses given event $E_p(d_s, d_p, d_e)$: $T_{E_p}^{Wiki}$, $T_{E_p}^{Twitter}$ and $T_{E_p}^{Weibo}$. To be specific, $T_{E_p}^{Wiki}(d)$, $T_{E_p}^{Twitter}(d)$ and $T_{E_p}^{Weibo}(d)$ denote the number of Wikipedia clicks, Twitter messages, and Weibo messages, respectively, related to E_p on date d. Figure 3 is a visual example of the three time-series vectors for *Obama Barack*.

3.4 Twitter and Weibo Comparison

The goal of this paper is to compare the differences in responses to a set of commonly interesting hot events shared between two distinct communities: Twitter and Weibo users. We propose difference indicators to compare both responses in three aspects: the degree of popularity, the temporal dynamic, and the information diffusion pattern. In the following statements, we use X to represent Twitter or Weibo community.

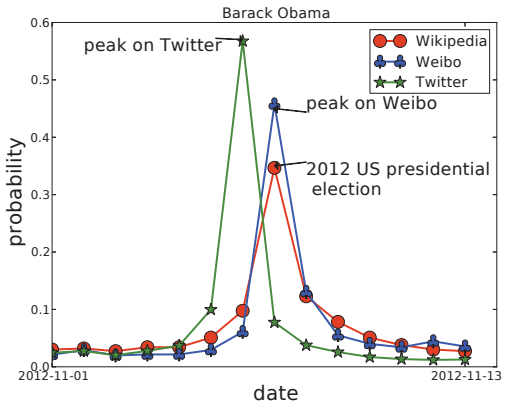


Figure 3: The Wikipedia, Twitter and Weibo response towards *Barack Obama*. All three vectors are normalized.

3.4.1 Popularity

The popularity of an event represents the degree of collective attention towards the event, which can be measured by the sum of daily probabilities that this event is mentioned on Twitter or Weibo during the time period of the event. To be specific, the popularity of event E_p is computed as:

$$P^X(E_p) = \sum_{d=d_s}^{d_e} \frac{T^X_{E_p}(d)}{N^X(d)} \quad (3)$$

where $T^X_{E_p}(d)$ is the number of messages in microblogging platform X related to E_p on d, and $N^X(d)$ is the total number of messages posted on d. We can further obtain the categorical popularity $P^X(C)$ for category C by averaging the values of $P(E_p)$ for all $E_p \in C$.

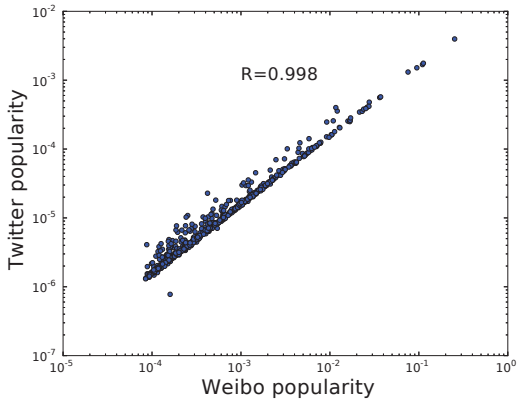


Figure 4: The correlation between Weibo popularity and Twitter popularity. The two popularity scores are highly correlated

3.4.2 Temporal Dynamic

We use d_p to denote the peak date of event E_p detected from Wikipedia click data. We are interested in when the discussion of E_p reaches the maximum degree on Twitter and Weibo and how the spiky discussion date is temporally related to d_p. Therefore, we compute the peak temporal delay as:

$$Delta^X(E_p) = d^X_p - d_p \quad (4)$$

where d^X_p denotes the dates when the volume of messages reaches the maximum on Weibo or Twitter. Similarly we can obtain the categorical peak delay $Delta^X(C)$ for category C by averaging the values of $Delta^X(E_p)$ for all $E_p \in C$.

To better understand the temporal dynamic between $T^{Twitter}_{E_p}$ and $T^{Weibo}_{E_p}$, we utilize KL- divergence, which is a non-symmetric measure of the difference between two probability distributions. After normalizing $T^{Twitter}_{E_p}$, $T^{Weibo}_{E_p}$ and $T^{Wiki}_{E_p}$, we calculate the KL-divergence as:

$$D^X_{kl}(T^X_{E_p}||T^{Wiki}_{E_p}) = \sum_{d=d_s}^{d_e} ln\frac{T^X_{E_p}(d)}{T^{Wiki}_{E_p}(d)} * T^X_{E_p}(d) \quad (5)$$

Again, we obtain the categorical KL-divergence $D^X_{kl}(C)$ for category C by averaging the value of $D^X_{kl}(T^X_{E_p}||T^{Wiki}_{E_p})$ for all $E_p \in C$.

3.4.3 Information Diffusion Pattern

Microblogging users access information via two types of ways: propagation through social network (internal information diffusion) or exposure to other channels (external information infection) [14]. In this study, we compare the dynamics of information diffusion in the social network environments of Weibo and Twitter. Specifically, we are trying to answer the following research question: Given an event E_p, what is the probability that a user is infected (i.e. discuss E_p) before the event peak d_p, denoted as $Pr^X(E_p, K)$, given that a number of neighbors (i.e. K neighbors) of the user in the social network, have already mentioned this event on the start date d_s of this event?

To construct social networks, we collect a large number of Weibo and Twitter messages and extract three types of relationships: forwarding, replying, and mentioning. For example, if u_i forwards to, replies to, or mentions u_j more than t times in historical data, we create a directed edge $u_i \rightarrow u_j$ on the social network.

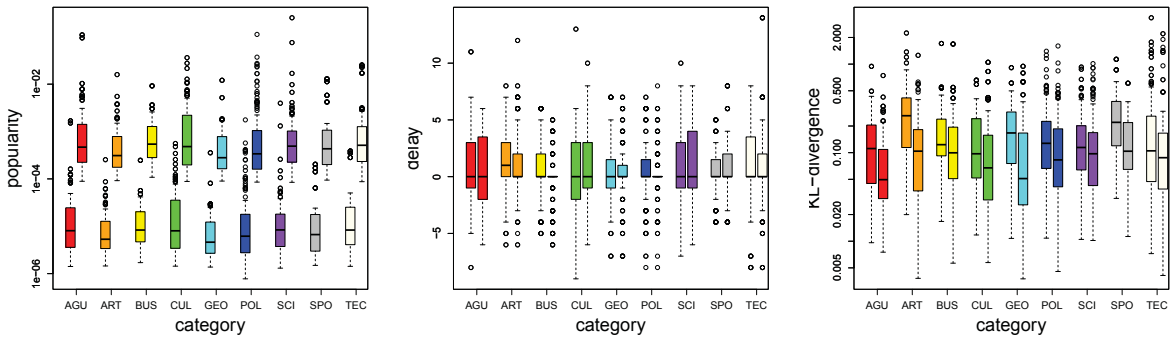
Consequently, we obtain a directed social network $G(V, E)$ where V is the set of users and E the set of edges $(u_i \rightarrow u_j)$ indicating that u_i previously forwarded, replied to, or mentioned u_j. Given an event E_p and value K, we define a *K-diffusion network*, which is a subgraph $G'_{E_p}(V', E')$, where $V' \subset V$ and $E' \subset E$. Specifically, $V' = S \cup N$ where S is the set of nodes in V representing the users who initiate the discussion of E_p before E_p starts (also called "seed users set"), and N is the set of nodes in V who are directly linked to K nodes in S; and $E' = \{(u'_i, u'_j)|u'_i \in \text{N}, u'_j \in \text{S}, u'_i \rightarrow u'_j\}$. We compute the diffusion probability $Pr^X(E_p, K)$ on $G'(V', E')$ as the fraction of users in N who also discussed E_p, after E_p starts until reaching the peak. (A similar method is introduced in [17].)

Again, we obtain the categorical diffusion probability curve $Pr(C, K)$ for category C by averaging the values of $Pr^X(E_p, K)$ for all $E_p \in C$.

4. EXPERIMENT

4.1 Data preparation

We dump a total of 3.4 million Wikipedia article pages (in the English version) and collect their click log statistics from Oct 15, 2012 to Nov 15, 2012. Then we rank those pages based on their aggregated daily click numbers during that time period and select the top 1% (i.e., 3.4 thousand) pages. We drop the other 99% pages since we only care about extremely hot events in our papers. After



(a) (b) (c)

Figure 5: The distribution of (a) popularity, (b) delay, and (c) KL-divergence for nine categories in Weibo and Twitter. For each category, the left box represents the distribution for Weibo while the right box represents the distribution for Twitter. Overall, Weibo and Twitter are similar in terms of relative popularity ranking but different in terms of temporal dynamic patterns.

applying the peak detection algorithm in Equation 1 (we empirically set the value of α to four) to the 3.4 thousand pages, around six thousand pages with daily click peaks are detected. After filtering out the pages without Chinese version, over three thousand pages with both Chinese and English version are left.

Next, we query the title of those Wikipedia pages in both Twitter and Weibo using methods proposed in Section 3. Based on the retrieval results, those pages with more than 50 hits returned from both Twitter and Weibo search index are selected. Furthermore, we classify these hot events into one of the 25 top categories defined in Wikipedia and pick the nine most representative categories. The category names, sample Wikipedia pages, as well as the page counts in each of the nine categories, are listed in Table 2.

In addition, to conduct the information diffusion analysis, we build up $G^X(V, E)$ by collecting the user's forwarding, replying and mentioning relationships (the threshold of communication frequency t is set to one) from Sep 15, 2012 to Oct 15, 2012, on both Twitter and Weibo. Consequently, $G^{Twitter}(V, E)$ contains 28 million nodes and 140 million edges, while $G^{Weibo}(V, E)$ contains 1 million nodes and 3.8 million edges. We can see that the size of $G^{Weibo}(V, E)$ is much smaller than $G^{Twitter}(V, E)$. Moreover, to construct seed users set S, for each event E_p, we search users who mentioned E_p during the period of $[d_s - 3, d_s - 1]$. Here we pick up a three-day pre-start period to find seed users with regards to E_p and we will investigate how to better select the period in our future work.

4.2 Popularity

After calculating the popularity scores for all hot events discussed in Weibo and Twitter, we compute the Pearson's correlation between Weibo popularity and Twitter popularity. Figure 4 shows that these two popularity scores are highly correlated, indicating that the degree of popularity of the selected hot events are very similar in Twitter and Weibo. In other words, if some event is popular in Twitter, it is likely that the same event is also popular in Weibo, vise versa.

The distribution of popularity scores in each category is characterized in Figure 5(a). Overall, those events are more popular in Twitter than Weibo, for all categories. It is because that those detected hot events are more favorable to Twitter user than Weibo users, as Twitter users are more likely to use Wikipedia than Weibo users (in China, the most popular online encyclopedia is Baidu Baike[2]). However, with regards to the relative categorical popularity, both Weibo and Twitter share almost the same ranking, again demonstrating that the event popularity scores are highly corrected between Weibo and Twitter. In particular, for both Weibo and Twitter, users' interests are more focused on *Agriculture*, *Business*, and *Culture* categories, while *Arts* and *Geography* categories are less popular.

4.3 Temporal Dynamic

The temporal dynamic of an event characterizes when and how users' response to the event reaches a peak. Particularly, *peak delay* indicates how fast the peak is reached, while *KL-divergence* depicts how diverse the pulses are between Weibo and Twitter. In either metrics, the pulse of the click log of Wikipedia servers as the baseline community response to certain event.

4.3.1 Peak delay

Figure 5(b) demonstrates the distribution of peak delay for all categories. Overall, both Weibo and Twitter users respond to hot events very fast (although Twitter response seems faster), with less than one day of delay on average. In particular, in *Politics* and *Business*, the delay is almost negligible on Twitter compared to Weibo.

Figure 3 is a case from *Politics* to illustrate the peak delay difference between Weibo and Twitter. Obviously, both Chinese and Americans show enormous attention towards the status of *Obama* during the 2012 US presidential election. However, the discussion about *Obama* reached its maximum on Twitter only one day after the election was held on Nov 6, 2012. By contrast, Weibo discussion and Wikipedia clicks about *Obama* reached the maximum one day behind the Twitter peak day. This demonstrates the potential use of Twitter as a predictive tool for political elections. Figure 6(a) shows another example from *Business*. The *Tommy Hilfiger* Corporation announced its deals for Black Friday 2012 around Oct 22, 2012, but the related discussion on Twitter reached its peak as early as almost one week before the announcement. Another interesting observation is the spot of the Weibo peak after the announcement day, even though it seems that Chinese users are irrelevant to Black Friday. The signal may be generated by some Chinese economists who are interested in American markets.

[2]http://baike.baidu.com/

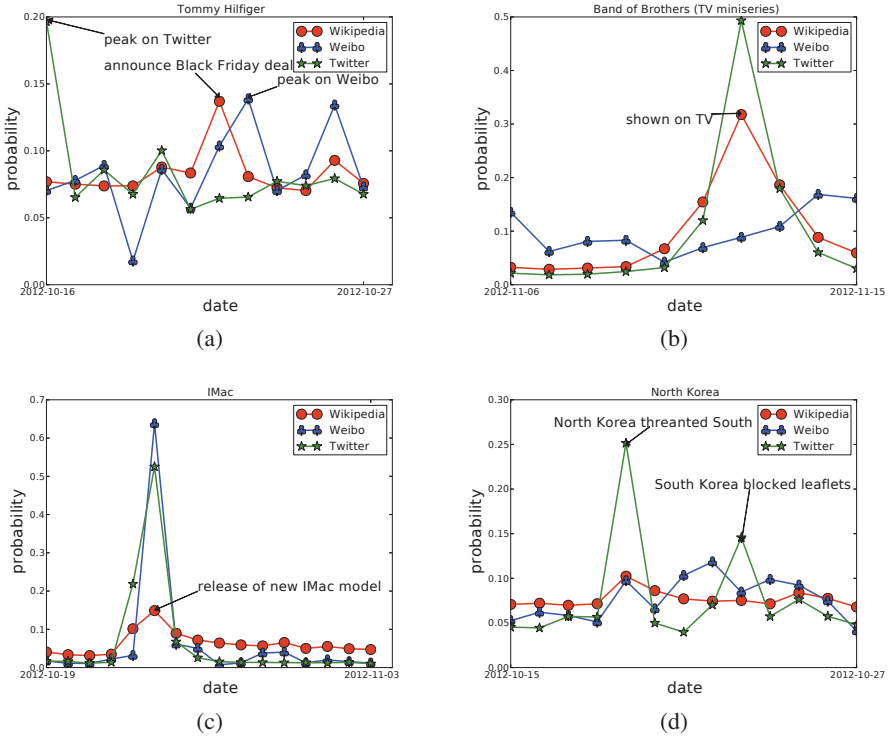


Figure 6: The temporal dynamic of users response curves shown for (a) Tommy Hilfiger (*Business*), (b) Band of Brothers (*Arts*) (c) IMac (*Technology*) and (d) North Korea (*Politics*). All are typical examples to illustrate the similarity ((c)) and differences ((a),(b),(d)) between Weibo and Twitter.

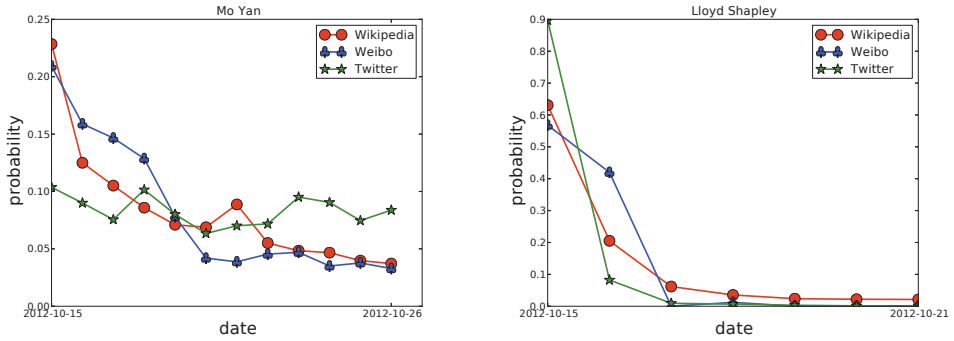
4.3.2 KL-divergence

Peak delay analysis only gives a one-point comparison (i.e. peak day) in the temporal response. To better investigate the difference in the whole process of users responses between Weibo and Twitter, we utilize KL-diverence to compare the overall temporal pulses between Weibo and Twitter during the time period of some event. Figure 5(c) shows KL-divergence distribution for all categories. Overall, the Weibo pulse follows Wikipedia pulse more closely than Twitter, implying that Twitter users' response is closer to Wikipedia users' response. Again, it is because that Twitter users are more likely to use Wikipedia than Weibo users.

Nevertheless, the degrees of difference in KL-divergence are not in the same level for all categories. In particular, Weibo and Twitter users' responses exhibit large gaps in *Arts*, *Culture* and *Geography*, while relatively similar trends in *Science* and *Technology* comparing with other categories. The reason of the categorical difference, we hypothesize, is that China and the Western world are quite different in cultural background and geographical location, but the people from the two communities have almost equal chance to get access to scientific and technological information through the Internet.

To better illustrate the temporal dynamic differences between Weibo and Twitter, we particularly select several instances from multiple categories. Figure 6(b) shows the users responses to *Band of Brothers* (belonging to *Arts*), a very popular TV show in the US. When it was shown on Spike TV on November 12, 2012, hot discussion was triggered on Twitter but not on Weibo, although this TV show is also well-known in China. Figure 6(c) demonstrates temporal pulses of *iMac* (belonging to *Technology*) when Apple released its new iMac model on October 23, 2012. We can see that Weibo and Twitter users respond to the technology news in an almost synchronized manner, with only small differences in the degree of peakiness. Figure 6(d) shows the difference between Weibo and Twitter users responses to the news about North Korea threatening South over propaganda balloons on Oct 19, 2012 by Reuters. The relationship between North and South Korea is always a political focus for both China and the the Western world. We can see that Twitter users respond much more intensely to this event than Weibo users. Moreover, there's even another Twitter spike spotted when South Korea claimed to block leaflets from the North on Oct 23, 2012. The difference between China and the Western world with regards to their political backgrounds and attitudes towards North and South Korea, may affect the media coverage and opinions about the same event, thus leading to different public attentions.

Finally, we list another two closely related examples to further illustrate the users responses difference between Weibo and Twitter from different categories. Figure 7 shows the collective attention towards the 2012 Nobel prize winner *Mo Yan* in literature (belonging to *Arts*) and *Lloyd Shapley* in economics (belonging to *Science*). When the announcement of *Mo Yan* winning the 2012 Nobel prize in literature was released on Oct 11, 2012, it was breaking news in China, because he is the first recognized Chinese person who ever won Nobel prize. Correspondingly, there's an obvious spike in Weibo response around Oct 11, 2012 (Figure 7(a) does not clearly show this spike since we cut off our data on Oct 15, 2012) and gradually the spiky discussion faded out. Such a spiky trend is not clearly seen on Twitter, even though people outside of China also paid attention to *Mo Yan* and kept steady attention. On the contrary, the news that American economist *Lloyd Shapley* won the



(a) (b)

Figure 7: The difference between Weibo and Twitter response towards 2012 Nobel prize winners (a) *Mo Yan* in literature (belonging to *Arts*) and (b) *Lloyd Shapley* in economics (belonging to *Science*). Weibo users paid more attention to *Mo Yan*'s news in literature while Weibo and Twitter users respond similarly to *Lloyd Shapley*'s news.

Nobel prize in Economic Sciences on Oct 15, 2012, aroused both spiky discussion on Twitter and Weibo users around that day, and then both trends went down sharply.

4.4 Information Diffusion Pattern

We have already compared the users' responses in Weibo and Twitter in terms of popularity and temporal dynamic, now we focus on how the underlying social networks structure can affect the users response. To be specific, we investigate whether the probability that a user will respond to some hot event (i.e. diffusion probability) is affected by the number of the user's neighbors who have already responded to the same event,

Figure 8 depicts the diffusion probability curve against different K on Twitter network for four different categories of events: *Arts*, *Politics*, *Science*, and *Technology*. Overall, the diffusion probability increases consistently when K increases (except for the *Technology* category, which stops increasing after some K value), which indicates a complex contagion phenomenon. In other words, a user is more contagious to the infection of an hot event when more of his/her neighbors have already responded to it earlier. Therefore, the social network in Twitter does play a role to facilitate information diffusion around those selected hot events. Moreover, there exist differences in the diffusion patterns among various categories. Specifically, the trend increase in *Arts* and *Politics* is faster than *Science* and *Technology* (while K increases), which implies that Twitter users are more likely to be affected by their neighbors by the infection events in *Arts* and *Politics* categories than *Science* and *Technology* categories. Social media users tend to be subjective about evens in *Arts* and *Politics* and more easily to be emotionally influenced by their neighbors; On the contrast, users generally hold objective opinions about events in *Science* and *Technology*, therefor less likely to be affected by their neighbors.

By contrast, on Weibo, we find the diffusion probability is significantly lower than that on Twitter, for all the categories. We can interpret this finding in two ways. First, compared with Twitter, Weibo users are less likely to be contagious to the target infection hot event via internal social networks (they might choose the external channels, i.e., news and other websites, to access the hot event). Second, as mentioned before, Twitter users are more likely to use Wikipedia than Weibo users. Specifically, those hot events sampled from Wikipedia, are not necessarily the 'global tast'. For instance, we find a large number of American musicians, artists, politicians, and athletes in the hot events extracted from Wikipedia query log, of whom Chinese Weibo users may have limited knowledge. This could be the reason that Weibo users are less likely to be contagious via the Weibo network.

Meanwhile, for the Weibo network, it is rare to find the diffusion phenomenon through social networks, when K is larger than three. We plot distribution of diffusion probability for all categories on Weibo in Figure 8(b) when $K = 1, 2, 3$. We can see that there's a huge increase in the diffusion probability when K increases from 1 to 2. Specifically, except for *Technology*, the diffusion probability increases more than twice over all other categories. Especially, in *Arts* and *Business*, the probability increases almost more than ten times more. There's a dip when K changes from 2 to 3, possibly due to the data sparsity problem. When $K > 3$, we hardly find events that diffuse through the Weibo network. This finding verifies our earlier assumption that Weibo users access (Wikipedia) hot events more randomly, instead of through Weibo social network.

5. CONCLUSION AND FUTURE WORK

In this paper, we compared how users of each respond to external hot events Twitter and Weibo, the two largest social media communities in the world. We employed Wikipedia as the research vehicle for hot event discovery, categorical aggregation, and English→Chinese translation, and proposed a set of static and dynamic indicators to compare Twitter and Weibo from three perspectives: popularity, temporal dynamic, and information diffusion. We want to share the following observations, drawn from our extensive evaluation on Twitter and Weibo data:

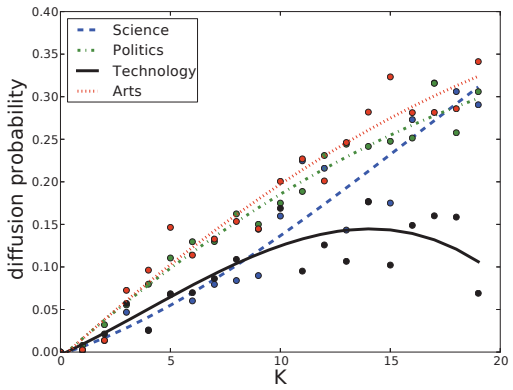
1. Based on the study of event popularity, we observed that Weibo and Twitter users share similar degree of interests towards a set of commonly interesting events between the two communities.

2. Based on the study of peak delay, we observed that both Weibo and Twitter users respond quickly to hot events, with less than one day delay from the peak of Wiki searches for most events. In addition, we observed that Twitter users respond faster than Weibo users on *Politics* and *Business* events. Our heuristics is that this is due to the ideological difference of the user base and/or information blockade in Mainland China.

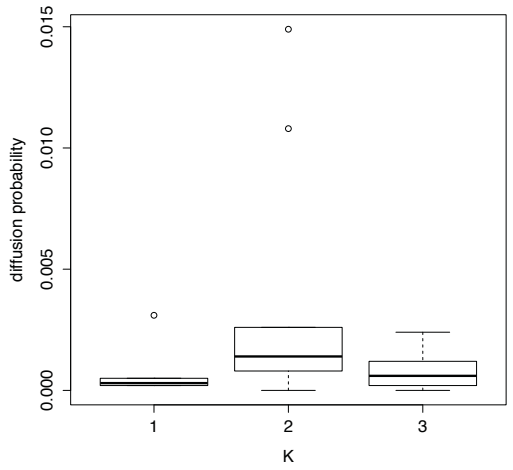
3. Based on the study of KL-divergence, we observed that Twitter and Wikipedia temporal pulses are relatively similar com-

paring with Twitter, which we attribute to the significant overlapping between Twitter and Wikipedia communities. We also observed that Weibo and Twitter users demonstrate very similar responses to *Science* and *Technology* events while the two communities' responses to events in *Arts*, *Sports*, *Politics* and *Culture* display larger gaps.

4. Based on the study of information diffusion, we observed that while Twitter users are more likely to be infected by hot events via Twitter social network (internal exposure), Weibo users access hot events more likely via other channels (external exposure).



(a)



(b)

Figure 8: The diffusion pattern of hot events through social networks in (a) Twitter and (b) Weibo. Specifically, (a) shows the diffusion probability curves of four categories and (b) shows the diffusion probability distribution when K=1,2,3

While interesting observations were discovered in the study described in this paper, the results are far from conclusive. We are aware of two important limitations of this study: (1) the language variation in Chinese, such as the use of homophone and metaphor to discuss certain hot (sensitive) events negatively affected the search in Weibo corpus, yielding fewer results than desired; (2) the purposive sampling method used in this study, e.g. using categorical cross-language event metadata offered by Wikipedia, may lead to bias as the overlapping between Twitter and Wikipedia communities is significantly larger than that between the Weibo and Wikipedia communities.

Our immediate next step is to address the limitations listed above. To address limitation (1), we will use more sophisticated text mining and natural language processing algorithms to find the latent semantic match results, instead of just focusing on explicit word search (statistical match). To address limitation (2), we will distinguish events that are *global* and *regional*, and compare Twitter and Weibo on these two types of events separately.

In summary, this is a pilot study that opens the door for more in-depth analysis of the social phenomenon boosted by blogging and instant messaging. We are looking forward to working with social scientist to further analyze the results for more insightful discoveries.

6. ACKNOWLEDGMENTS

Xin Shuai thanks the National Science Foundation for its support of his PhD research under grant SBE #0914939; Tian Xia thanks the Beijing Higher Education Young Elite Teacher Project for its support of his work.

7. REFERENCES

[1] Eric Baucom, Azade Sanjari, Xiaozhong Liu, and Miao Chen. Mirroring the real world in social media: twitter, geolocation, and sentiment analysis. In *Proceedings of the 2013 international workshop on Mining unstructured big data using natural language processing*, pages 61–68. ACM, 2013.

[2] Johan Bollen, Huina Mao, and Alberto Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *The International AAAI Conference on Weblogs and Social Media*, 2011.

[3] Surajit Chaudhuri, Raghu Ramakrishnan, and Gerhard Weikum. Integrating db and ir technologies: What is the sound of one hand clapping? In *Conference on Innovative Data Systems Research*, pages 1–12, 2005.

[4] Qi Gao, Fabian Abel, Geert-Jan Houben, and Yong Yu. A comparative study of users' microblogging behavior on sina weibo and twitter. In *User Modeling, Adaptation, and Personalization*, pages 88–101. Springer, 2012.

[5] Wanqiu Guan, Haoyu Gao, Mingmin Yang, Yuan Li, Haixin Ma, Weining Qian, Zhigang Cao, and Xiaoguang Yang. Analyzing user behavior of the micro-blogging website sina weibo during hot social events. *Physica A: Statistical Mechanics and its Applications*, 395:340–351, 2014.

[6] Bernard J Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. Twitter power: Tweets as electronic word of mouth. *Journal of the American society for information science and technology*, 60(11):2169–2188, 2009.

[7] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.

[8] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[9] David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. English stop word list, 2004. [Online; accessed 2014-02-07].

[10] Daifeng Li, Jingwei Zhang, Gordon Guo-zheng Sun, Jie Tang, Ying Ding, and Zhipeng Luo. What is the nature of chinese microblogging: Unveiling the unique features of tencent weibo. *arXiv preprint arXiv:1211.2197*, 2012.

[11] Xiaozhong Liu and Howard Turtle. Real-time user interest modeling for real-time ranking. *Journal of the American Society for Information Science and Technology*, 64(8):1557–1576, 2013.

[12] Xiaozhong Liu and Vadim von Brzeski. Computational community interest for ranking. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 245–254. ACM, 2009.

[13] Evgeny Morozov. *The net delusion: The dark side of Internet freedom*. PublicAffairs Store, 2012.

[14] Seth A Myers, Chenguang Zhu, and Jure Leskovec. Information diffusion and external influence in networks. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 33–41. ACM, 2012.

[15] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *The International Conference on Language Resources and Evaluation*, 2010.

[16] Joshua Ritterman, Miles Osborne, and Ewan Klein. Using prediction markets and twitter to predict a swine flu pandemic. In *1st international workshop on mining social media*, 2009.

[17] Daniel M Romero, Brendan Meeder, and Jon Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 695–704. ACM, 2011.

[18] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

[19] Xin Shuai, Xiaozhong Liu, and Johan Bollen. Improving news ranking by community tweets. In *Proceedings of the 21st international conference companion on World Wide Web*, pages 1227–1232. ACM, 2012.

[20] Jian Sun. Ansj-seg chinese segmenter. [Online; accessed 2014-02-07].

[21] Sherry Turkle. *Alone together: Why we expect more from technology and less from each other*. Basic Books, 2012.

[22] Sebastián Valenzuela, Namsu Park, and Kerk F Kee. Is there social capital in a social network site?: Facebook use and college students' life satisfaction, trust, and participation1. *Journal of Computer-Mediated Communication*, 14(4):875–901, 2009.

[23] Wikipedia. List of websites blocked in china, 2014. [Online; accessed 2014-02-07].

[24] Fareed Zakaria. *The post-American world: release 2.0*. WW Norton & Company, 2011.

[25] Bin Zhao, Zhao Zhang, Yanhui Gu, Xueqing Gong, Weining Qian, and Aoying Zhou. Discovering collective viewpoints on micro-blogging events based on community and temporal aspects. In *Advanced Data Mining and Applications*, pages 270–284. Springer, 2011.