Selected Papers of Internet Research 16:
The 16th Annual Meeting of the
Association of Internet Researchers
Phoenix, AZ, USA / 21-24 October 2015

THE CHALLENGES OF WEIBO FOR DATA-DRIVEN DIGITAL MEDIA RESEARCH

Jing Zeng
Queensland University of Technology

Jean Burgess
Queensland University of Technology

Axel Bruns
Queensland University of Technology

Introduction

Data generated via user activity on social media platforms is routinely used for research across a wide range of social sciences and humanities disciplines. The availability of data through the Twitter APIs in particular has enabled new modes of research, including in media and communication studies; however, there are practical and political issues with gaining access to such data, and with the consequences of how that access is controlled.

In their chapter 'Easy Data, Hard Data', Burgess and Bruns (2015) discuss both the practical and political aspects of Twitter data as they relate to academic research, describing how communication research has been enabled, shaped and constrained by Twitter's 'regimes of access' to data, the politics of data use, and the emerging economies of data exchange. This conceptual model, including the 'easy data, hard data' formulation, can also be applied to Sina Weibo.

This paper builds on this model to explore the practical and political challenges and opportunities associated with the 'regimes of access' to Weibo data, and their consequences for digital media and communication studies. We argue that in the case of Weibo and in the Chinese context, the politics of data access can be even more complicated than in the case of Twitter, which makes academic scholarship relying on 'big social data' (Burgess & Bruns, 2012) from this platform more challenging.

Weibo and Weibo studies

The word *weibo* means 'micro-blog' in Chinese, and therefore weibo is often considered to be 'like Twitter, only for China'. There are a number of micro-blogging platforms in China, but this paper will focus on Sina Weibo (hereafter to referred to as Weibo),

Suggested Citation (APA): Zeng, J., Burgess, J., & Bruns, A. (2015, October 21-24) *The Challenges Of Weibo For Data-Driven Digital Media Research.* Paper presented at Internet Research 16: The 16th Annual Meeting of the Association of Internet Researchers. Phoenix, AZ, USA: AoIR. Retrieved from http://spir.aoir.org.

launched in 2009 by one of China's biggest Internet portals, Sina Corp. Weibo has more than 500 million registered users and 61 million daily active users (CNNIC 2014). As the larger social and media ecology in China changes, the role of Weibo in Chinese society and the wider Chinese-speaking diaspora continues to evolve. In its early years, Weibo was largely used by people in mainland China to share everyday or 'trivial' content (Yu et al. 2011: 8). Mirroring the parallel legitimization or 'de-banalisation' process that Twitter went through (Rogers, 2014), from 2012 the Weibo platform was visibly associated with various forms of user-led activism, leading observers to speculate on the platform's political potential in addition to its more mundane, everyday uses (Hassic, 2012; Chan *et. al.,* 2012). Since 2013, as the result of the Chinese Communist Party's dramatic escalation of its efforts to tighten its control over Weibo-based communication (Guo & Ying, 2015), as well as the competition from other social media services in mainland China, Chinese internet users are no longer as active in micro-blogging platforms as they previously were (CNCC, 2014). However, its apparently declining influence within China does nothing to diminish Weibo's significance as an object of study – indeed its history and evolving social role make it an even more dynamic site for digital media research, but the current state of the art in Weibo scholarship leaves much to be desired.

Compared to Twitter, Weibo scholarship in general is still at a preliminary stage of development, with notable limits on both the diversity of topics and the standardization of methods. Much of the work on Weibo in international journals is dominated by institutions outside mainland China: language barriers continue to deter mainland scholars from publishing in English. As a result of this lack of connectivity and knowledge exchange, western Weibo scholarship is very concerned with political issues understood 'from the outside', particularly those related to the possible 'democratizing' potential of Weibo in relation to assumed democratic deficits associated with the Chinese State and its censorship regime (e.g. King *et al*. 2014; Zhu *et al*. 2013; Ng & Landry 2013). These are important issues (albeit at risk of oversimplification), but the sociocultural role of Weibo and the range of uses to which it is put are far richer and more diverse than such approaches would suggest.

More Chinese research on Weibo made more accessible to international scholars would be beneficial to internet studies in general, since Chinese microblogging and other everyday digital media use is such an important part of the global ecology of social media, and not only 'in China'. Likewise, Western scholars wishing to explore the applicability of hypotheses and findings in the context of Chinese microblogging need to develop a better sense of the platform's structure, affordances, and potential for data-driven research. As part of this movement towards more globalized digital media and communication research, this paper sketches out the possibilities and challenges of translating methodologies from Twitter research to Weibo. As a modest first step, it focuses on comparing the academic affordances of Twitter data and Weibo data.

Technical challenges

The Application Programming Interfaces (APIs) of major social media platforms are designed primarily to enable third-party developers to invent new applications of and for the platforms in question, within terms set by and aligned with the business models of the companies that own and provide them. They are also crucial tools for researchers to

access social media data – they are therefore important *mediators* of digital media research.

Twitter offers three types of APIs: 1. the Streaming API, which provides a continuous stream of new posts matching criteria set by the researcher; 2. the Search API, which can retrieve a limited number of historical tweets matching given criteria (often used in combination with the Streaming API); and 3. the REST API, which provides programmatic access to author new posts and to read a user's profile and timeline. Although Weibo's APIs do not use the same typology, in terms of functionality they were very similar to those offered by Twitter, at least at first. From 2012, Weibo began to commercialize its APIs, however, and it has increasingly enforced restrictions on data crawling, imposing both IP-based and account-based rate limiting. These strategies are also widely used by other social media platforms, but Weibo's restrictions are among the most severe. Additionally, it introduced fees for the use of some of its APIs. Researchers requiring access to large sets of Weibo data will need to pay either Weibo itself or one of its 19 licensed data resellers (Weibo, 2015). Only one such company provides Weibo data to researchers outside China (*ibid*.). As a result of these monetization strategies, only a limited number of Weibo API functions can still be accessed for free by academic researchers. We now turn to provide further explanation of what Weibo APIs can and cannot do, with comparison to Twitter APIs.

While Twitter's Streaming API is the most widely used data collection source for Twitter scholars (Gaffney and Puschmann, 2014), in the case of Weibo, researchers can query Weibo's public timeline API to obtain the latest available posts. However, while Twitter's Streaming API keeps pushing posts to the user as new content becomes available, a single request of Weibo's public timeline API returns a maximum of only 200 posts. To retrieve a 'stream' of (presumably) real time posts from Weibo, multiple requests are needed. Given Weibo's increasingly tough restrictions on data-crawling, this process is highly inefficient and cumbersome.

Across platforms, search APIs are widely used when studying micro-blog-based discussion on a specific event or topic. However, Weibo's Search API is no longer available for the public. For the same task that Twitter's Search API can accomplish, Weibo researchers may need to reverse-engineer their own search mechanisms from scratch. First, they need to generate a long list of user IDs, encompassing hundreds of thousands (e.g. Fu *et. al.* 2013) and sometimes even millions of individual accounts (Guo *et. al.,* 2013). Second, they need to capture all of these accounts' posts, in order to generate a large, topic-independent corpus of user-generated content. Finally, they are then able to conduct keyword searches on this self-generated corpus of posts. One could use the user-timeline API to access numerous users' timelines, but now Weibo only allows its API users to crawl their own timelines.

An alternative way to crawl other users' timelines on a large scale is through crawling the researcher's own friends' updates. This is to say, researchers may need to set up a number of 'dummy' accounts, and use these accounts to follow target users. Most data-driven Weibo research requires the ongoing addition of new people to follow in order to enlarge the sample size. Earlier, this could be done through Weibo's add-friend API. Unfortunately, the add-friend function is now also no longer freely available. Now one may have to use some other more technically demanding mechanism to complete this

task. For example, researchers could build automated tools to add friends to their accounts through the platform's web interface.

This example demonstrates that, while Weibo continually tightens up its API access, it still remains technically possible to get around the obstacles. However, the consequence is an increasingly high technical barrier for data collection, and a lock-out of those researchers who have limited financial and technical support. This story is not unique to Weibo, because the Twitter API is becoming increasingly commercialized and restrictive as well (Burgess & Bruns, 2015). The difference, however, is in the frequency and transparency of how the two platforms change their rules of the game.

Geopolitical challenges

When studying Weibo, it is also essential to take the geopolitical context of this platform into consideration, because the Communist Party of China has had an active, interventionist role in Weibo's launch, boom, and recent decline. In 2009, after a number of violent riots in China's Uyghur region, the Chinese government shut down most social media sites operating in the country, including Twitter and Facebook. This became a direct factor leading to the birth of made-for-China social media sites: only one month after Twitter had been blocked in China, Sina Corp launched its micro-blogging product Weibo, which quickly became the most popular social media site in the country. The relationship it has with Chinese authorities (Benney, 2014) has been an important factor contributing to Weibo's success. From the very beginning, for instance, it worked closely with the state to censor information on the site. However, the political atmosphere around China's Internet has changed since Xi Jingping came into power. The new administration has taken a series of measures to 'clean up' China's Internet[1]. As some early Weibo studies indicate, the chilling effect of such regulations is real and measurable (Fu & Chaun, 2013; Ford, 2015).

The political atmosphere in China not only restricts academic freedom in terms of what kind of Weibo research can be done from inside China, but also impacts on what data can be obtained. For example, there is a long and still growing list of banned search terms on Weibo (Qiang, 2011; King *et al*., 2013). Academic data collection can be adversely affected by the complexity and uncertainty of such censoring practices on Weibo. For studies related to politically sensitive topics, the effect of content censorship is most direct. For example, in 2012 the state shut down the comments feature on Weibo for several days in an attempt to suppress information about the Bo Xilai incident. A more recent case are the Occupy Hong Kong protests. During these protests, Beijing tried to censor all non-official information on Weibo that related to the events. According to a censorship monitoring project at the University of Hong Kong, WeiboScope[2], the rate of posts removed from the system reached its 2014 peak during this period.

1. Since 2012, users are required to register their real identity with the service. From 2013, a five-strikes-and-out rule will see anyone posting five tweets on "sensitive" subjects have their account suspended. From the same year, Weibo users can be jailed for posting false information if it is forwarded more than 500 times or viewed over 5,000 times.

2. For details of this project, please refers to Fu, K., Chan, C., & Chau, M. (2013). Assessing censorship on microblogs in China: Discriminatory keyword analysis and the real-name registration policy. *Internet Computing, IEEE* 17(3): 42-50.

Data integrity challenges

Data integrity problems arise in the first place from the Weibo APIs' sampling methods. There is no documentation concerning the percentage of full public posts that Weibo's public timeline API provides (in the case of Twitter's streaming API, we know that the total volume of results returned is limited to one percent of the total current volume of global Twitter activity at that moment). Zhu *et al.*'s (2012) study estimates that 1.7 to 10 percent of the full public timeline data become available through Weibo's public timeline API. Since this estimate is rather broad and there is limited scholarship exploring this issue, we still know very little about how much data we can actually retrieve from Weibo's public timeline API. The second problem with Weibo's public timeline API is that the data is not randomly selected. Because there is a few minutes' delay between the time a post is published and the time it becomes available on the public timeline, it is suspected that the data on the public timeline may have been manipulated. For instance Zhu *et al.*'s (2012) study suggests that censorship has been applied over what can be displayed on Weibo's public timeline, as it is found that certain posts with sensitive keywords do not appear on the public timeline but are still shown on users' individual timelines. This lack of knowledge regarding the Weibo APIs' sampling methods can have severe impacts on the reliability and generalizability of studies. Of course the issue of sample bias is not unique to research drawing on Weibo data. Researchers have found evidence showing that Twitter's 'random' streaming API has biases as well. However, the body of Twitter scholarship generally indicates that efforts have been made to identify sources of possible bias in the sampling process. For instance, Morstatter (2013) finds that Twitter's Streaming API appears to have a bias toward tweets that may have more value for research. By contrast, in the case of Weibo, there is little scholarship dealing with this issue apart from examining the impact of content censorship on Weibo's API sampling. More comprehensively exploring the issue of data integrity with respect to sampling is therefore an opportunity for Weibo researchers.

Spam is a further issue affecting data integrity in Weibo research. In its the quest for advertising revenues, Weibo Corp shapes the user experience design of the platform towards consumption and entertainment, which in turn relies on user growth and visitor numbers. Perhaps because of this, the company has demonstrated little commitment to removing the large proportion of spam and fake accounts on the platform (Yu *et al*., 2012; Cheng *et al.,* 2013; Fu & Chau, 2013). This "artificial inflation" (Yu *et al*., 2012) of user activity on the platform is closely aligned with the company's interests, especially since it launched its IPO in the U.S. in 2014. For scholars, the very high proportion of spam on Weibo complicates the conclusions about public communication that can be drawn from Weibo data. Cheng *et al.*'s (2013) research indicates that certain types of spam can be detected with algorithms that examine users' following patterns; on the other hand, spam is simply part of the communicative ecology of Weibo and it may not always be desirable to 'clean' it from the sample. Finding accessible ways of identifying and, if necessary, removing spam is a further opportunity for Weibo researchers.

Conclusion

Exploring China's social media landscape is essential for developing a better understanding of the global media ecology. Therefore, the discussion of Weibo's potential should not be restricted to its hypothetical democratizing effect, but be better grounded in openly available, data-driven research drawing on a combination of

rigorous and reliable, quantitative, qualitative and critical approaches. This article has discussed how 'hard' access to and research-oriented usage of Weibo data can be compared to Twitter – but 'hard data' and 'easy data' are, of course, relative terms. Even though there are increasingly high technical and economic barriers of access to Weibo data, Weibo is still relatively open and accessible compared to other Chinese media. As it is one of a handful of windows through which we can spot the dynamics of Chinese social media platforms – which collectively host the world's biggest online population –, an important next step is to develop the means for institutions or researchers with access to large-scale data to make their data, methods and code available to other researchers, without putting themselves or Weibo participants at risk. Existing examples of such initiatives include the work of Fu and Chan (2013) and Ding et al. (2013), projects which have made their data available for scholars in the field.

References

Benney, J. (2014). The aesthetics of Chinese microblogging: State and market control of Weibo. *Asiascape: Digital Asia, 1*(3), 169-200. Retrieved from http://booksandjournals.brillonline.com/content/journals/10.1163/22142312-12340011.

Burgess, J., & Bruns, A. (2012). Twitter archives and the challenges of 'big social data' for media and communication research. *M/C Journal* 15(5)**.** Retrieved from http://www.journal.media-culture.org.au/index.php/mcjournal/article/viewArticle/561

Burgess, J., & Bruns, A. (2015). Easy data, hard data: The politics and pragmatics of Twitter research after the computational turn. In G. Langlois, J. Redden, & G. Elmer (Eds.),*Compromised data: From social media to big data.* Bloomsbury, London. (In Press)

Chan, M., Wu, X., Hao, Y., Xi, R., & Jin, T. (2012). Microblogging, online expression, and political efficacy among young Chinese citizens: The moderating role of information and entertainment needs in the use of Weibo. *Cyberpsychology, Behavior, and Social Networking, 15*(7), 345-349. Retrieved from http://online.liebertpub.com/doi/pdf/10.1089/cyber.2012.0109.

Cheng, B., Fu, J., & Zhen, M. (2013). Detecting zombie in Sina microblog: A machine learning approach. *International Journal of Advancements in Computing Technology,* 5 (2): 612-620. Retrieved from http://www.aicit.org/IJACT/ppl/IJACT2070PPL.pdf.

China Internet Network Information Center (CNNIC). 2014. Statistical survey report on the internet development in China. Retrieved from http://www1.cnnic.cn/IDR/ReportDownloads/201404/U020140417607531610855.pdf

Ding, C., Chen, Y., & Fu, X. (2013). Crowd crawling: towards collaborative data collection for large-scale online social networks. *Proceedings of the first ACM conference on online social networks, Boston, Massachusetts, USA*.

Ford, C. A. (2015). *China looks at the west: Identity, global ambitions, and the future of Sino-American relations*. Lexington: University Press of Kentucky.

Fu, K., Chan, C., & Chau, M. (2013). Assessing censorship on microblogs in China: Discriminatory keyword analysis and the real-name registration policy. *Internet Computing, IEEE* 17(3): 42-50. Retrieved from http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=6459495

Fu, K., & Chau, M. (2013). Reality check for the Chinese microblog space: A random sampling approach. *PloS One* 8(3). Retrieved from http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0058356

Gaffney, D., & Puschmann, C. (2014). Data collection on Twitter. In K. Weller, A. Bruns, J. Burgess, C. Puschmann, & M. Mahrt (Eds.), *Twitter and Society* (pp. 55-68). New York: Peter Lang.

Guo, Z., Huang, J., He, J., Hei, X., & Wu, D. (2013). Unveiling the patterns of video tweeting: A Sina Weibo-based measurement study. In M. Roughan & R. Chang (Eds.), *Passive and Active Measurement* (Vol. 7799, pp. 166-175): Springer Berlin Heidelberg.

Guo, B., & Jiang, Y. (2015). Analyzing the coexistence of emerging transparency and tight political control on Weibo. *The Journal of International Communication, 21*(1), 78-108. Retrieved from http://dx.doi.org/10.1080/13216597.2014.998700.

Hassid, J. (2012). The politics of China's emerging micro-blogs: Something new or more of the same? APSA 2012 Annual Meeting Paper. Retrieved from http://ssrn.com/abstract=2106459.

King, G., Pan, J. & Roberts, M. E. (2013). How censorship in China allows government criticism but silences collective expression. *American Political Science Review, 107*(02), 326-343. Retrieved from http://dx.doi.org/10.1017/S0003055413000014.

King, G., Pan, J., & Roberts, M. E. (2014). Reverse-engineering censorship in China: Randomized experimentation and participant observation. *Science,* 345(6199). Retrieved from http://www.sciencemag.org/content/345/6199/1251722.

Li, Y., Gao, H., Yang, M., Guan, W., Ma, H., Qian, W. &Yang, X. (2013). What are Chinese talking about in hot Weibos? *arXiv*: 1304.4682: 546-557. Retrieved from http://arxiv.org/pdf/1304.4682.pdf.

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. *arXiv*: 1306.5204. Retrieved from http://arxiv.org/abs/1306.5204.

Ng, J. Q., & Landry, P.F. (2013). The political hierarchy of censorship: An analysis of keyword blocking of CCP officials' names on Sina Weibo before and after the 2012 National Congress (s)election (June 15, 2013). Eleventh Chinese Internet Research Conference, 2013, Forthcoming. Retrieved from http://ssrn.com/abstract=2267367.

Qiang, X. (2011). The battle for the Chinese internet. *Journal of Democracy, 22*(2), 47-61. Retrieved from http://muse.jhu.edu/journals/jod/summary/v022/22.2.xiao.html.

Rogers, R. (2014). Debanalising Twitter: The transformation of an object of study. In K. Weller, A. Bruns, J. Burgess, C. Puschmann, & M. Mahrt (Eds.), *Twitter and Society* (pp. ix-xxvi). New York: Peter Lang.

Weibo (2015) List of data trader partners. Retrieved from http://open.weibo.com/wp/partner.

Yu, L., Asur, S., & Huberman, B. A. (2011). What trends in Chinese social media. *arXiv*: 1107.3522. Retrieved from http://www.hpl.hp.com/research/scl/papers/chinatrends/china_trends.pdf

Yu, L., Asur, S., & Huberman, B. A. (2012). Artificial inflation: The true story of trends in Sina Weibo. Paper presented at the 5th SNA-KDD Workshop '11, San Diego, Calif. Retrieved from http://arxiv.org/abs/1202.0327.

Zhu, T, Phipps, D., Pridgen, A., Crandall, J. R, & Wallach, D. S. (2012). Tracking and quantifying censorship on a Chinese microblogging site. *arXiv*: 1211.6166. Retrieved from http://arxiv.org/abs/1211.6166.

Zhu, T., Phipps, D., Pridgen, A., Crandall, J.R., & Wallach, D.S. (2013). The Velocity of censorship: High-fidelity detection of microblog post deletions. *arXiv*: 1303.0597. Retrieved from http://arxiv.org/vc/arxiv/papers/1303/1303.0597v1.pdf.