Policy & Internet
Vol. X, No. X, Month 200X, 000–000

Towards Area-Smart Data Science: Critical Questions for Working with Big Data from China

Daniela Stockmann[a] [*]

Hertie School of Governance, Berlin

Abstract: While the Internet was created without much governmental oversight, gradually states have drawn territorial borders via Internet governance. China stands out as a promoter of such a territorial-based approach. China's separate Web infrastructure shapes data when information technologies capture traces of human behavior. As a result, area-expertise can contribute to the substantive, methodological, and ethical debates surrounding big data. In this article, I discuss how a number of critical questions that have been raised about big data more generally apply to the Chinese context: How does big data change our understanding of China? What are the limitations of big data from China? What is the context in which big data is generated in China? Who has access to big data and who knows the tools? How can big data from China be used in an ethical way? These questions are meant to spark conversations about best practices for collaboration between data scientists and China experts.

Keywords: China, social media, big data, research methods, ethics

[a] Please direct correspondence to: Daniela Stockmann, Professor of Digital Politics and Media, Hertie School of Governance, Friedrichstrasse 180, 10117 Berlin, Germany. Email: stockmann@hertie-school.org

[*] The research leading to the results on new media has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n. [338478]. The Hertie School of Governance and Leiden University are both beneficiaries of the grant. For more information on this project, entitled "Authoritarianism 2.0: The Internet, Political Discussion, and Authoritarian Rule in China," see www.authoritarianism.net.

Introduction

Nation states have returned as key players in governing the digital world. In response to September 11, 2001 many governments have strengthened their abilities to take advantage of online user data in an attempt to protect their citizens. After Edward Snowden revealed the US government's massive data collection through the National Security agency in 2013, many countries have legitimized similar initiatives by passing new national security laws that increased surveillance and censorship within their territorial borders.[1] Terrorist attacks in Europe and the United States have increased pressure on Google, Facebook, Twitter, and other technological companies to cooperate more closely with law enforcement, creating different patterns of content across borders (Hintz, 2016). Of course, Internet governance has also been previously carried out by national governments, Internet service providers (ISPs) and other actors (van Eeten and Mueller 2012), but recent developments suggest that governments are tightening control over information technology within territorial borders.

China stands out as a country that has promoted regulation of the Internet within national borders for a long time by establishing a separate Web infrastructure that remains "Chinese." Besides constructing the well-known Chinese firewall, the central government has ensured that China's national Internet backbone remains in the hands of Chinese companies (Jiang 2013, Pan 2017). In response to Google's spat with China over censorship, cyber attacks and Google's license renewal to operate in China in 2010, the central government has promoted Internet sovereignty encouraging individual countries to independently choose their own path of cyber

[1] 14 out of 65 countries passed such laws in 2015 and in 2016. See Freedom House (2015) "Freedom of the Net 2015," available at https://freedomhouse.org/report/freedom-net/freedom-net-2015; Freedom House (2016) "Freedom of the Net 2016," available at https://freedomhouse.org/report/freedom-net/freedom-net-2016.

development and regulation while participating in international cyberspace governance (Jiang 2010).[2]

The "Internet Sovereignty" approach of Internet governance that China favors and advocates holds much sway in the Global South where many countries are authoritarian. At the 2012 World Conference on International Telecommunications 89 countries out of 144 (including China and Russia, and many Arab countries) voted in favor of nation-state based regulation of the Internet, while 55 (including the US, EU member states, India, and Japan) voted against it (Jiang 2013). While the early Internet was created without a great deal of governmental oversight, gradually "borders have been drawn around the previously borderless forms of cyberspace" (Hintz 2016, p. 328).

Despite this growth of territorial borders in Internet governance, area studies have been remarkably absent from the debate about issues of big data.[3] This is surprising since big data is often created when information technologies are capturing traces of human behavior. Individuals often use multiple communication devices that can capture and distribute text, images, audios, and videos. Traces can be marked with coordinates of time and place, creating continuous records of activity. Buildings, vehicles, and public places are instrumented with similar technologies (Borgman 2015). Laws, regulations, policies, initiatives, and infrastructures governing information technology shape data. As states take a strong role in governing

[2] Internet sovereignty is not a Chinese invention: France has applied the concept of national sovereignty to the debate about transnational data flows in the 1970s. Drake, W. J. 1993. "Territoriality and Intangibility: Transborder Data Flows and National Sovereignty." *In:* Nordenstreng, K. and H. I. Schiller eds. *Beyond National Sovereignty: International Communications in the 1990s.* Norwood: Ablex, 259-313. Since the first World Internet Conference in Wuzhen in 2014, China has promoted the concept internationally. See *Xinhua*. 2015. "China Voice: The World needs fresh rules for Internet Governance." February 9, 2017. http://news.xinhuanet.com/english/2015-12/18/c_134930950.htm.

[3] The only reference to Chinese studies I have come across is King, G. 2014. "Restructuring the Social Sciences: Reflections from Harvard's Institute for Quantitative Social Science." *PS: Political Science & Politics,* 47(01), 165-172.

information technology, area expertise may prove useful when mining, analyzing, and sharing big data.

In this article, I take the perspective of area-smart data science. Area specialists are often (mistakenly) placed in the same category as qualitative and ethnographic researchers, who have been outspoken about the relationship between these research methods and big data research (see, for example, Kozinets 2010). In contrast to the common belief that area expertise is mainly based on knowledge of language or factual information about context and events, I do not reduce area studies to contextual knowledge. Area expertise is an interdisciplinary approach that brings together scholars with diverse disciplinary and methodological backgrounds who have made a long-term commitment to a specific area or region in the world (Katzenstein 2001, O'Brien 2011, Pieke 2013). As such, area-studies also encompass an understanding of theoretical and methodological debates relevant to a specific geographical location in the world.

The reflections regarding area-smart data science grew out of a co-organized project as part of the Digital Methods Initiative at the University of Amsterdam (DMI) aimed at training professors, students, and professionals in digital methods.[4] My focus here is mainly on big data and digital methods as they apply to social media. My observations regarding China studies and big data were developed based on systematic analysis of peer-reviewed articles in influential social science and area studies journals. Conversations with colleagues who self-identify as China scholars based in East Asian Studies, Sinology, Political Science, Communication Studies, Sociology, and Public Policy also inform my thinking. Building on the thoughts developed in this paper, together with Blake Miller and Jessica Batke, I have co-founded ChinaDataLab.org as a

[4] I am a political scientist and communication scholar focusing on China who has recently started to experiment with the use of Chinese social media data. Trained at the University of Michigan, my approach to scholarly research places the research question at the center of scientific inquiry: in my own work, I usually combine qualitative with quantitative research methods, choosing the method appropriate to answer the question at hand.

platform where China observers can collaborate, share data, develop technical expertise, and publish relevant findings.

Here, my aim is to discuss how a number of critical questions raised by boyd and Crawford (2012) apply to big data from China. As a main promoter of Internet sovereignty China is an extreme case, but similar area-based understandings may also prove useful when other countries or regions that are increasingly moving towards an approach to governing information technology based on territorial borders.

The first question discussed below relates to the value of big data for scientific inquiry. As scholars are debating the impact of the data revolution on the production of knowledge, many researchers have argued against popular accounts of big data as a replacement of the scientific method (see, for example, Kitchin 2013, Clark and Golder 2015). One prominent example constitutes Lazer et al.'s study demonstrating that Google Flu Trends did not produce the valid indicators of flue incidences that it initially appeared to be. As the authors put it: "quantity of data does not mean that one can ignore foundational issues of measurement and construct validity and reliability and dependencies among data" (Lazer et al. 2014). Of course, big data is not only defined by its size, but by a number of key features (see, for example, Dodge and Kitchin 2005, Marz and Warren 2012, boyd and Crawford 2012, Mayer-Schönberger and Cukier 2013). It also poses exciting new opportunities for scholarly inquiry (see, for example, Kitchin 2013, Monroe et al. 2015). Yet in the context of China - a traditionally data-scarce area of research that is increasingly drawing attention due to China's increasing importance in the world - one may wonder what insights we gain from these new data sources and emerging methods. Data about China has become more easily accessible, larger, and more varied over time, and researchers now rely more on methods that are unique to the Internet. As sources and methods to assess knowledge change, so may be scientific understandings of China. Unfortunately, a systematic review of journal articles leaves one rather disappointed about the current state of big data

research in the China field: new data sources and methods did not seem to lead to new questions about China and have not taken advantage of the full potential of new data sources and emerging methods. Perhaps this is not surprising given the general slow uptake of many disciplines, but there is certainly more room for contributions that truly challenge existing paradigms or raise original questions that reshape existing debates or spark new ones.

In light of this discussion, I then pose two questions that highlight limitations of big data from China: the first question is related to data acquisition that challenges the common assumption that big data is cheap. Sudden and rapidly changing circumstances often interrupt access to data and pose major challenges for comparison over time, the ability to replicate, and research planning. Due to the lack of scholarly discussion on such data acquisition challenges currently, successful examples of how these difficulties have been overcome remain potentially underappreciated by China scholars not engaging in this type of research or by data scientists working with data from outside the Great Firewall of China. The second question also highlights limitations of data and bias, but as a result of the context within which big data from China is generated. Most importantly, the Chinese state's active manipulation and management of the Internet creates biases in data that are used to draw conclusions about China. We also lack knowledge about the characteristics of active users, bots, spammers, and paid bloggers who create social media content. Such knowledge is needed in order to better assess what can be learned about China from big data.

A second set of questions emerges out of the practicalities of conducting big data research in the Chinese context. Within the Great Firewall, China has been remarkably successful in protecting its domestic technology companies, many of which generate big data, from competition with its mostly US-based counterparts (see, for example, Pan 2017). By separating itself from the global Internet, building technological and institutional means of control, and protecting domestic businesses, China aims to maintain the capacity to deal with the challenges of big data in a way

that stabilizes its rule (Jiang 2013, Pan 2017). This approach has implications for transnational data flows and research ethics. Scholars from inside and outside of China are in the same boat as they start to embrace new sources and methods, but they also have different levels of access to data and tools: on the one hand, China will produce an army of data scientists and domain researchers working together in areas that are in line with the broader visions and policies of the Chinese leadership; on the other hand, researchers outside of China have a competitive advantage when it comes to politically sensitive topics, but their data sources are also more limited to leaked or publicly accessible data. If big data research is conducted at the expense of fieldwork, there is the danger of losing touch with local communities. As researchers from abroad are trying to protect collaborators and participants, they may also find themselves in a dilemma to make a decision as to which regulatory regime they adhere to. Big data originating from or hosted inside the Great Firewall of China sometimes requires different approaches to research ethics than what is considered ethical in Europe or the United States.

In the conclusion, I tie the answers to each question to research collaboration between data scientists and domain researchers working on China. Such collaborations are needed to address the difficulties surrounding big data. I do not intend to suggest that big data generated in China necessarily differs from other regions. However, China has separated its network from the rest of the world, and these conditions have created specific characteristics of those data that require expertise on China to properly address them. An area-smart data science benefits from area expertise to take full advantage of this new and exciting data source.

How Does Big Data Change our Understanding of China?

Researchers interested in China used to struggle with lack of data. Facing information scarcity during the Cultural Revolution, China scholars often relied on a small number of sources providing snapshots about the country, such as media reports, migrants who had exited the system, or

Foreign Broadcast Information Service (FBIS), an open intelligence service by the CIA. The field was small with everyone knowing each other personally and sharing information in order to puzzle together a broader picture of China (O'Brien, 2011).

Today, the situation could not be more different. As with scholarship more broadly, digital methods are increasingly used in scholarship on China. As an example, I collected articles in the Science and Social Science Citation Indices that mention China and social media between 2010 and 2015.[5] Figure 1 displays how percentages of articles matching the keywords "China" and "social media" developed in comparison to articles only mentioning "China" or only "social media." Over time, interest in Chinese social media has grown in scholarship across disciplines. Among articles mentioning "China" and "social media," digital methods were frequently used to collect or analyze data: 63 percent of articles published in Science journals and 12 percent of articles in Social Science journals used digital methods. As shown in figure 2, in both kinds of journals, the trend for using digital methods is increasing. According to Rogers (2013), digital methods are specific to the virtual environment of the Internet, for example, by using crawling or APIs to collect data or using cross link or sentiment analysis to analyze data. Conventional social science methods, such as online experiments or online surveys, for example, are adapted to the online environment and called digitized methods: they rarely require application of computational techniques to search, aggregate, and cross-reference large data sets, often generated as digital traces by social interaction (Welser et al. 2008, boyd and Crawford 2012).

Insert Figures 1 and 2 about here

[5] Obviously, these articles can only provide us with rough estimates for cross-disciplinary differences. Computer Scientists publish predominantly in peer-reviewed conference papers, while publication of books and book chapters is common in Area Studies. These publications are not covered in Web of Science Citation Indices.

The increasing use of digital methods regarding Chinese social media derives in part from the new opportunities to study human behavior in areas that could previously not be examined by researchers. Schroeder describes these as "possibilities of making advances in understanding phenomena without necessarily controlling them in practice" (Schroeder 2014). For example, one area of research in Computer Science links the spatial locations with certain activities taken by people at different times of day (see, for example, Wu et al. 2014, Liu et al. 2014). Cheng and Manion (2015) take advantage of such geotagged microblog posts to link geographical location within 2 and 25 meters at the moment of posting with a political attitude expressed in the post. Prior to social media data this geographical precision was unheard of, especially when linked to public opinion and behavior. Created by businesses for commercial purposes researchers have little input in the creation of such data, but these sources can reveal information that was previously out of reach for scientists.

As promising as these new data sources are, at the moment they are rarely used to pose new questions about China. Instead, most studies provide insights on long-standing academic discussions by exploring public reactions to specific events (see, for example, Gu and Ye 2014, Gu et al. 2014, Wang et al. 2014, Fu and Chau 2014), investigating censorship and political control over Chinese media (see, for example, Fu et al. 2013, King et al. 2013, Auer and Fu 2015, Cairns and Carlson 2016), and detecting trends in public attention and emotions in China (see, for example, Li et al. 2014, Fan et al. 2014 2015). These existing studies have not taken full advantage of the innovative potential of digital methods to provide insights into aspects that computational social science is strong at, such as explaining how group behavior, norms, institutions and other social phenomena evolve or how social learning occurs (see, for example, Conte et al. 2012).

Out of these works, a new vision of Chinese society seems to be emerging. Big data research describes Chinese social media users as a dynamic, pluralist, and interactive whole – perhaps composed of different kinds of users assuming different roles, but nevertheless as a collective group.[6] Traditional social science research methods were developed without access to terabytes of data describing their minute-to-minute interactions and locations of entire populations and individuals. As such they tend to take 'snapshots' of specific points in time and place the individual at the center of analysis, usually aggregating to one group level, if at all. Computational social science research methods, on the other hand, aim to explain (or predict) the complexities of real socio-economic systems (see, for example, Lazer et al. 2009, Conte et al. 2012). This complex systems approach reveals how individuals interact with one another at various group levels (groups, networks, communities) and how they change their attitudes and behaviors in very short intervals of time at specific locations. Pentland (2014) describes such methods as "social physics" which use mathematical models to understand how ideas flow through networks. Online social network analysis, sentiment analysis, data mining, and other computation research methods have vastly expanded our understanding of the circumstances, characteristics of users, and messages that contribute or disrupt collective action and information flows traveling through social networks (see, for example, Niu et al. 2013, Wang et al. 2014). The specific circumstances and characteristics of users and messages that shape social change, evolution, and learning have yet to be studied more in depth for China. Emerging research draws attention to the potential networking power of women and people living in certain cities in China as well as the power of emotions associated with specific kinds of issues, such as the Diaoyu/Senkaku island dispute or food security (Fu and Chau 2014, Fan et al. 2014). As a result, the complexity and dynamic nature

[6] Of course, there are exceptions to these broader observations. See, for example, Nip, J. Y. M. and K.-w. Fu. 2016. "Challenging Official Propaganda? Public Opinion Leaders on Sina Weibo." *The China Quarterly,* 225, 122-144.

of Chinese society has become more evident to scholars, perhaps especially for those without extensive experience of living in China.

In the sample of journal articles I collected, a publication that used digital methods received, on average, about 3.4 more citations in comparison to publications that used conventional social science research methods.[7] This is a very crude estimate of influence in scholarship, but it illustrates that digital methods draw attention from peers publishing in journals with a general rather than area studies audience. Therefore, big data research may facilitate the adoption of China as a topic into the broader disciplinary discussions beyond Chinese studies.

Boyd and Crawford (2012) suspected that big data research leads to a focus on studying reactions to events of the present or immediate past because of the sheer impossibility or difficulty to access older data. With respect to China, big data research is unlikely to have such an effect. In light of China's rapid transformation, scholars of contemporary China have faced a similar situation for a long time. In order to avoid publications being outdated before publication, researchers face pressure to constantly update and often focus on short-term developments regardless of the research methods used. Furthermore, in the digital humanities, big data research has improved our understanding of traditional rather than modern China (see, for example, de Weerdt In Press, Mostern 2011, Garnaut 2014). Instead, big data research may help to visualize the complex and dynamic nature of Chinese society, and allow this vision to travel beyond the China field.

After these observations regarding the value of big data regarding scientific inquiry I now turn to limitations.

[7] Articles using traditional social science methods received, on average 1.8 citations (s.d. 2.9); articles using digital methods received, on average, 5.2 citations (s.d. 9.9). Citations data refers to total citations as of January 2016.

What are the Limitations of Big Data Acquisition from China?

Much of the attention devoted to big data research originates in its scale that is different from what was available before (Schroeder 2014). Such power in numbers adds empirical evidence to arguments that are difficult to achieve with other types of evidence. Big data research on China - just as big data research more generally - rarely explains biases or limitations of data as a result of data acquisition. The relevance of data quality "is somehow overshadowed by the sheer volume of information that we now have at our disposal" (González-Bailón 2013).

Social scientists often stress that the size of data is meaningless if the characteristics of the sample are unknown (see, for example, boyd and Crawford 2012, González-Bailón 2013, Vis 2013). Because computers operate based on logical processes, methods of big data collection can produce bias in data sets that may be unintended, undesirable, or nonsensical. To quote a widely cited expression from computer science, "garbage in, garbage out."

From the articles on Chinese social media sampled earlier, data acquisition predominantly relied on application programming interface (API) provided by the company that owns relevant data. Data stream is sampled based on a streaming algorithm, which has limited memory and processing time per item. As a result, an API works like a filter in a data river: it never returns the full historical data, but always filters out data with certain characteristics cleared for external access by the company. Data scientists often prefer to acquire data through APIs, because the data is returned in a data-interchange format (JSON) that is easy to read and write. [8]

It is important to understand that social media data are usually business data. As such, the data are owned by private companies that restrict third parties to gain access to user data. Some Chinese social media companies have followed the example of Twitter and Facebook to open up

[8] Conversations with product managers and data scientists at Wire, Research Gate, Twitter, Google, Sina, Baidu, and Tencent.

certain parts of their data to the public. IT companies do so because they often acquire other innovative technologies rather than building new products, platforms, and interfaces internally; apart from giving developers access, they may also provide a certain amount of data as a token of public service. [9] However, APIs are not aimed at generating representative data and a sample drawn from them always depends on the specific features of the API and time when sample was drawn (see, for example, Gerlitz and Rieder 2013, González-Bailón et al. 2014, Morstatter et al. 2013, Tromble and Stockmann 2017).

In my sample of articles on Chinese social media, the vast majority drew data from Sina Weibo, a popular Twitter-like social media platform. Underrepresentation of other social media platforms may be a function of difficulties to access other social media data or the importance that Weibo was given in studies of Chinese social media before the rise of WeChat as its main competitor around 2011.[10] Sina Weibo provides two sets of APIs: the developer APIs, which are obtained with explicit consent of the company, and the open APIs.[11] Most Weibo research is based on these open APIs with some noticeable exceptions (see, for example, Chen et al. 2013). Like Twitter and other social media, the specific algorithm to sample the data is not publicly known, and it therefore remains unclear which tweets are left out in a sample obtained through the open API.

One challenge in Weibo research is the fact that Sina changes the algorithm of open APIs frequently for reasons unclear to outsiders. As a result, temporal comparisons can only be made

[9] Conversations with product managers and data scientists at Wire, Research Gate, Twitter, Google, Sina, Baidu, and Tencent.
[10] Conversations with product managers and data scientists at Wire, Research Gate, Twitter, Google, Sina, Baidu, and Tencent.
[11] For a description of Sina Weibo's open APIs in Chinese see open.weibo.com/wiki/%E5%BE%AE%E5%8D%9AAPI, accessed December 15, 2015.

within periods of time within which the open API remained stable.[12] In contrast to Twitter where researchers have time to discuss the implications of different APIs for the representativeness of the sample (González-Bailón et al. 2014, Vis 2013, Nagler and Tucker 2015), by the time of publication insights into characteristics of Sina Weibo's APIs are often outdated. For example, Fu and Chau's (2013) exemplary study of Weibo users drawing a random sample from user accounts has changed substantially since and randomization has become more difficult, according to the authors. Of course, Twitter APIs are not entirely transparent, but more continuity allows for more time to evaluate their use for research (Liang and Fu 2015, Tromble and Stockmann 2017).

To facilitate data acquisition for those not trained in data science, third-party developers have started to provide access to data acquired through APIs (or other means). One third-party developer often used by researchers on China is DiscoverText, a cloud-based software developed by Political Scientist Stuart Shulmann, which enables data collection and data mining of Sina Weibo, Tencent Weibo as well as Twitter and other non-Chinese social media.[13] DiscoverText accesses Weibo through Boardreader, a company that promises access to the largest possible stream, but it remains unclear how the data is acquired and which API is used. While Discovertext and other third-party developers acknowledge the limitations of data collection, these may be hard to trace for researchers when using intermediary platforms, especially if there is little awareness or considerations regarding them (see also Vis 2013).

Difficulties of data acquisition associated with China's rapid transformation is not new to China scholars, of course, but it constitutes a heightened challenge in the Chinese context, posing limitations on comparisons over time, the ability to replicate, and planning of future research projects. As big data from China becomes more influential in scholarly discussion, more space

[12] This is a major reason why Weiboscope has so far only been able to make certain "chunks" of data available for further research. Thanks to Fu King-wa for sharing this information.
[13] www. http://discovertext.com/, accessed March 23, 2015.

needs to be devoted to explaining how data was obtained and how the resulting characteristics of the sample are linked to research questions and conclusions. To address difficulties of data acquisition, it is helpful to involve a data scientist early on and to be prepared to spend considerable time on having to adapt to unforeseen changes in technologies during data acquisition.

In addition, the context within which data are created also poses limitations.

What is the Context in which Big Data is Generated in China?

Understanding how data are created is key to making the data useful for scientific discovery (Borgman 2015). Taken out of context, data loses its meaning (boyd and Crawford 2012). As we interpret big data from China, we know much about certain aspects that shape its creation, but less about others.

It is well known that the Chinese state has built a vast infrastructure to control, monitor, and shape digital information sources. While increased regulation is not specific to authoritarian regimes (Morozov 2011, MacKinnon 2012), especially politically closed states like China are channeling resources into censoring digital technologies and creating regulatory environments that shape the structure of big data. Since the emergence of the Chinese Internet in the 1990s, China has built an extensive system for Internet control, surveillance, and manipulation. This system includes configuration of Internet gateway infrastructure (Boas 2006), blocking websites and filtering (Chase and Mulvenon 2002), Internet policing (Brady 2008), regulation of Internet service providers (MacKinnon 2009), suppression of dissident use and discipline of cyber cafes (Chase and Mulvenon 2002, Qiu 2000), and employment of web commentators to shape and alter public debate (Bandurski 2008, Han 2015a, Miller 2016, King et al. In Press). In comparison to his predecessor Hu Jintao, Xi Jinping has made “informatization” of Chinese society and “cybersecurity” a political priority, releasing a set of programs and measures, such as for example,

the creation of a national Cyberspace Administration in 2013 (Alsabah 2016). As part of these efforts the Chinese Communist Party aims to move Chinese society into the digital age while also preventing potentially negative consequences via increased censorship, surveillance, and management of the Internet. Big data and data science are playing a key role in this endeavor. One of the most prominent examples is a social credit system which integrates data originally used for advertisement or other commercial reasons like online shopping data, social media data, and data on mobility collected via apps or smartphones. Companies like Alibaba, Tencent or Baidu develop this system and provide data for the state (Wang 2015). While implementation of such a system is underway, integration of such massive amounts of data and infrastructures is not only technically challenging but also politically sensitive as access to data also potentially reshapes the power distribution between different levels of administration (Meissner and Wübbeke 2016).

Besides the well-known Great Firewall of China, these regulatory practices, institutional and technological infrastructures separate digital China from the global Internet: people living in mainland China are part of a separate digital environment than those living in Taiwan, Hong Kong, and abroad (see, for example, Schneider 2015). While it makes sense to separate big data from mainland China from other regulatory regimes, in practice Chinese Internet control is more fragmented and decentralized than commonly assumed. In an environment where censorship is operationalized at multiple levels of government and in partnership between numerous institutions, technology and industry, researchers cannot assume that the process of censorship is nationally standardized (Greitens 2013). Research on censorship in social media points towards bias against information that criticizes political leaders (Gallagher and Miller 2017), has the potential to mobilize collective action (King et al. 2013, King et al. 2014), and against certain regions, including Tibet, Qinghai, and Ningxia, and Beijing (Bamman et al. 2012, Wright 2014). However, these results depend on the method of censorship studied and can change over time

(see, for example Wright et al. 2015). Little is known about how specific censorship practices relate to bias in data from China.

Another major limitation of data results from its lack of representativeness in comparison to the population of China. Social media data are created by active users. According to the China Network Information Center, the number of Internet users has grown to approximately 731 million in December 2016, and this growth has been taking place largely in China's most developed coastal provinces and urban areas. With the rise of smart phones social media users have rapidly grown, but their actual numbers and characteristics in comparison to Internet users are uncertain. WeChat claims 500 million monthly active users, Weibo 175.7 million monthly active users; and Baidu Tieba has an average of 50 million new posts per day.[14] Just as in other contexts, some of these accounts are bots that produce automated content without directly involving a person; some users have multiple accounts, while some accounts are used by multiple people (see also boyd and Crawford 2012). One unique aspect of Chinese cyberspace are the large numbers of spammers and inactive accounts (Yu et al. 2012) and paid bloggers, the so-called 50 cent party members, who actively produce pro-government messages (Bandurski 2008, Han 2015b, Miller 2016, King et al. In Press). In order to better interpret social media data, we need to better understand the characteristics of users and other actors that generate the data.[15]

Next I turn to the practicalities of using big data from China, discussing access and research ethics.

[14] On Wechat see http://tech.sina.com.cn/i/2015-03-18/doc-icczmvun6903718.shtml, accessed on 28 May 2015. On Sina Weibo see http://tech.sina.com.cn/i/2015-03-11/doc-iawzuney0631454.shtml, accessed on 28 May 2015. Baidu Tieba's users count is based on an Interview with a Baidu Tieba senior manager (72906), 15 Apr 2015.

[15] Miller has developed a promising technique to identify paid bloggers in Chinese cyberspace, which could be used as a starting point to differentiate between these groups Miller, B. A. P. 2016. "Automatic Detection of Comment Propaganda in Chinese Media." [online]. https://papers.ssrn.com/sol3/Papers.cfm?abstractid=2738325.

Who has Access to Big Data and Who Knows the Tools?

The scale of big data suddenly makes studies that used to be considered "large n" appear small. "In the world of computational social science, Big Data has provoked an analytic arms race to work with more data, better data, bigger data in pursuit of discovering so-called truths about the social world. […] ''How big is your dataset?'' they ask. ''1500,'' I say, ''no bigger than a modest survey, but different in an important way," writes Faucault Welles (2014), a junior communication scholar, about the need to justify the size of data in her research. Previously perceived "quantoids" who work primarily with quantitative data may suddenly find themselves in a position that qualitative researchers have been in for some time. Big data research challenges our scholarly identities that have been manifested in scholarship over a long period of time.

As such, the trend towards digital methods provides an opportunity for scholars of developing countries like China to catch up with the methodological skill-set of researchers trained in Europe and the US. In the humanities and social sciences, the application of data science is still in its infancy. Scholars trained outside and inside China are in the same boat as researchers are only starting to learn how to best make use of big data.

On the one hand, scholars based in China may be better equipped to obtain access to big data and learn how to analyze big data efficiently. China's plans for using big data will produce a large demand for data scientists and data analysts, who know the context that explains the patterns in the data.[16] Large international ICT firms that try to get their foot in the door of the

[16] Under Xi Jinping plans have been developed to increasingly use information technology as a political instrument for surveillance. Implementation of social credit scoring remains challenging Meissner, M. and J. Wübbeke. 2016. "IT-backed Authoritarianism: Information Technology Enhances Central Authority and Control Capacity under Xi Jinping." *In:* Heilmann, S. and M. Stepan eds. *China's Core Executive: Leadership Styles, Structures and Processes under Xi Jinping.* 52-57.

Chinese big data market, such as IBM, for example, provide help in training of this army of data scientists.[17] However, who gets access to training and data and at which organizations, companies, and universities remains to be seen. Boyd and Crawford (2012) suspect that those with money or those inside technology companies have advantages at getting access to big data. In China, personal connections provide additional opportunities for outsiders to gain access to business data, and being inside China provides an advantage to establish those networks.[18]

The infrastructures supporting big data research in China are also actively shaped by the political goals of the Chinese leadership, especially Xi Jinping's vision for informatization and cybersecurity of the Chinese society. In 2015 the State Council officially announced the development of big data as a national strategy to improve business and technological innovation as well as governance and surveillance (Zeng 2016). Big data is used to address problems of traffic and transportation of China's large population (Ran 2013); China is also building a massive e-governance and cybersecurity system where the application of big data is less well understood (Schlaeger 2013, Schlaeger and Jiang 2014). In which policy areas the application of big data is encouraged and for which purpose will strongly influence who will get access to data and how those data will be analyzed and interpreted. Access from outside China is likely to be discouraged rather than supported though this infrastructure.

On the other hand, publicly accessible data, such as Weibo and other social media, facilitate research from abroad without the necessity of spending extensive time in China. This allows research on topics that authoritarian governments tend to restrict, such as censorship, for example (Greitens 2013), and that are difficult to research from inside the system. Scholarship

[17] IBM Commits US$100 million to Support China to Nurture Big Data and Analytics Talent, https://www-03.ibm.com/press/us/en/pressrelease/44342.wss, accessed January 12, 2016.
[18] For example, in 2014 I was promised access to Sina weibo full data from a Sina employee; unfortunately, the person moved to another company, so I was unable to retrieve the data.

inside China may become technically more sophisticated, but only contribute to academic discussions considered to have practical applications to China's transformation. Certain voices may be discouraged by the Chinese leadership.

On sensitive topics, researchers from outside of China have a competitive advantage, but there is also the danger of interfering or losing touch with the very local communities that researchers claim to draw conclusions about. For example, King, Pan, and Roberts' (2014) influential study of censorship in China has been criticized for deceiving readers of manipulated content posted online and influencing results of researchers studying social media messages. King, Pan, and Roberts acknowledged such potential ethical concerns and pledged to "avoid, wherever possible, influencing or disturbing the system we are studying" (King et al, 2014, p. 1251722-9). In addition to interfering with local communities, greater access from abroad reduces time spent on fieldwork and immersing oneself in local communities and cultures. In the case of the Arab Spring, some local academics saw themselves reduced to "educed to mics saw themselves . In Chin who jet in and jet out" (Abaza 2011, p. 1). Of course, such kind of "academic tourism" is only one way to conduct research. Area studies has a long tradition of being aware of inequalities, and an area-smart data science could be designed in a way that shrinks the divide between scholarship in China and abroad instead of expanding it.[19]

How is Big Data from China used in an Ethical Way?

[19] Of course, area studies is not free from power relations between researchers based in China and abroad. For example, positivist standards for evaluation of research constitute an obstacle for researchers based in China to be published in journals based outside of China (conversation with editorial board members of an interdisciplinary Chinese studies journal, February 2017). However, an awareness of inequalities among China experts has sparked a lively discussion about how to address inequalities.

Ethics in big data research is mostly discussed with respect to concerns about privacy. In the case of social media, data are proprietary and users are not necessarily aware of all the uses, profit, and gains that come from the information they have posted. As a result, it has become problematic for researchers to justify their actions as ethical simply because data are accessible. Institutional Review Boards (IRBs) and other ethical review boards often ask whether human beings involved in the creation of data gave informed consent (boyd and Crawford 2012). Under the 2016 law for the protection of personal data by the European Commission, processing personal data for research on human beings within the European Union has to be conducted with explicit consent of participants, and researchers have to anonymize data before it is made public or ask for explicit consent of participants for such publication. According to the ethical standards of the European Commission, research conducted outside of the European Union, like China, should conform to the same ethical standards as requested in the European Union.[20]

These ethical standards assume that Chinese are just as much concerned about privacy and that explicit consent addresses these concerns. However, little is known about what concerns Chinese users have regarding privacy and whether explicit consent is the right way to address them. Empirical research on privacy in China often concentrate on exposure of personal data for fear of intrusion by government agencies (see, for example, Wu et al. 2011, Clinton et al. 2014).[21] Paradoxically, requests to give signatures or click a box on a website that is associated with an

[20] In my case, I was asked to retrieve explicit consent regarding the use of social media data in my research directly from participants. Based on ethical review process of my research project "Authoritarianism2.0," funded by the European Research Council. See also "Ethics for Researchers," European Commission, http://ec.europa.eu/research/participants/data/ref/fp7/89888/ethics-for-researchers_en.pdf, accessed January 12, 2016.

[21] Emerging research suggests that ordinary citizens may be more strongly concerned about social reputation as opposed to fear of political surveillance and censorship Stockmann, D. and T. Luo. In Press. "Which Social Media Facilitate Online Public Opinion in China?" *Problems of Post-Communism*..

IP address as proof of explicit informed consent can undermine rather than strengthen trust between participants and researchers. In societies where governments rule by law rather than being subject to rule of law signatures create a track record that may endanger rather than protect participants.[22]

The practical realities of data collection under the Chinese regulatory regime regarding privacy also provide a challenge to collect data anonymously. On part of the government, the Hu-Wen administration established a working group to draft a Personal Information Protection Act under the Hu-Wen administration, but generally online privacy enjoys only a modicum of legislative protection in China (Wu et al. 2011). User agreements of Chinese social media, such as Wechat, Sina Weibo, and Renren, ask for user consent to give access to data and give consent to deletion of data for political reasons (Stockmann and Zheng 2015). In some digital methods of data collection, social media technology requires the researcher to open a user account and to therefore give consent to user agreements. Researchers also have to make decisions about the online regulatory regime in which they intend to host data on servers. If users are required to jump the Great Firewall for data collection, there may be a trade-off between increasing representativeness of data by hosting data on servers inside the firewall and lowering the risk of exposure of personal data for political reasons by hosting data on servers in online regulatory regimes with high levels of legal online privacy protection.

These are some examples of ethical questions that may arise when big data is generated in China. When data is produced digitally, researchers are confronted with the instruments through which the Chinese state manages flows of information and data. Just like Internet users

[22] In this context, rule by law signifies societies in which law constitutes an instrument of governance, while rule of law signifies governance within the boundaries of the law.

within the Chinese online regulatory regime researchers also have to decide how to best address state control over data.

Towards Area-Smart Data Science

What makes big data research different from other methodological advances in the past in that it increasingly takes place in interdisciplinary collaborations (Wallach 2016). One individual is unlikely to be conversant in the substantive and methodological debates surrounding the use of big data in multiple research communities (Watts 2013). Even researchers with dual degrees in Computer Science and a disciplinary domain admit that they need to collaborate to gain access to data and equipment, learn about technical problem-solving, or knowledge about context (for example, language, region, and substantive issues).[23] Often, the scale and technical skills required to address challenges of big data research are more adequately addressed in larger research teams.[24]

This raises the question as to how such research collaboration with expertise on China would look like? Such an approach acknowledges that the Chinese context creates a number of China-specific issues that have consequences for evaluating the quality of scientific inquiry and research practices. These China-specific characteristics are self-imposed by the state's

[23] Conversations with Yunya Song at Hong Kong Baptist University or King-wa Fu at the University of Hong Kong, December 2015.

[24] On interdisciplinary collaboration see Klein, J. 2004. "Prospects for transdisciplinarity." *Futures,* 36(4), 515-526. Mullins, J. 2007. "Enabling International Access to Scientific Data Sets: Creation of the Distributed Data Curation Center (D2C2)." *28th Annual Conference of the International Association of Technological University Libraries (IATUL).* Stockholm: Libraries Research Publications. Ford, H. 2014. "Big Data and Small: Collaborations between ethnographers and data scientists." [online], 1 (2). February 8, 2017.

deliberate separation of its technological, institutional, and regulatory infrastructure surrounding big data from the rest of the world.

An area-smart data science involves scholarly discussions on substance, methodology, and research ethics. Systematic review of journal articles indicates that the substantive debate would benefit from China-specific knowledge to assess what scholarly insights derived from big data are original and innovative, but rely on data science to visualize the complex and dynamic nature of Chinese society and to allow this vision to better travel beyond the China field. Data science with greater area expertise will be of higher scientific value. With higher citation rates, substantive arguments built on data science will likely have a stronger impact on disciplinary debates.

In terms of methodology more research is needed regarding data bias. While much is known about the institutional and technological infrastructure and regulatory practices that shape the creation of big data in China, less is known about biases as a result of censorship practices, active users, bots, spammers, and trolls. Being conscientious about such biases requires more extensive explanations of data limitations in publications. For example, it is particularly important to explain how data were obtained and how data acquisition shaped the characteristics of the sample in light of China's rapid transformation.

A final discussion evolves around the standards and norms of ethical research practices. Different modes of access to data from inside and outside China may lead to further division between scholarship in China and abroad, but there is also the potential for scientists based in China to become more integrated in global scholarship. Inside and outside the Chinese firewall researchers are confronted with the instruments through which the Chinese state manages flows of information and have to develop strategies to cope with them in ways that protect research participants. Small interdisciplinary settings, such as workshops, may be a great way to bring researchers together, share best practices, and develop recommendations regarding how to deal with these issues for ethical review committees and journals.

Of course, many of these issues are also relevant to big data research more generally. Yet China's nation-state based approach towards governing information technology has created conditions that magnify the need to take into account bias and limitations of data while also being sensitive to inequalities and appropriate ethical research practices. And the Chinese case may not remain an exception: China's approach has a certain appeal to many governments in the Global South. And even democracies in Europe seem to increasingly turn towards tightening policies and regulations at the national and EU-level to address concerns about terrorism, privacy, and fake news.[25] This growth of territorial borders in Internet governance creates the need to take into account regional conditions that shape big data research. Big data with Chinese characteristics may only be the beginning.

[25] There is limited systematic empirical research on an overall trend across policy areas, but research on specific initiatives regarding counter-terrorism and privacy points towards this direction Argomaniz, J., O. Bures and C. Kaunert. 2015. "A Decade of EU Counter-Terrorism and Intelligence: A Critical Assessment." *Intelligence and National Security,* 30(2-3), 191-206, Gerry Qc, F. and N. Berova. 2014. "The rule of law online: Treating data like the sale of goods: Lessons for the internet from OECD and CISG and sacking Google as the regulator." *Computer Law & Security Review,* 30(5), 465-481. To address concerns about fake news Germany has passed a "Netwerkdurchsetzungsgesetz" in 2017, and similar initiatives are debated at the EU-level. See http://www.bundestag.de/dokumente/textarchiv/2017/kw26-de-netzwerkdurchsetzungsgesetz/513398, accessed July 18, 2017; See also http://www.europarl.europa.eu/news/en/press-room/20170329IPR69072/hate-speech-and-fake-news-remove-content-impose-fines-foster-media-literacy, accessed July 18, 2017.

Figures

Figure 1. Percentage of Articles on Chinese Social Media in Comparison to Articles on Social Media and on China. Source: Web of Science Citation Index, 2010-2015.

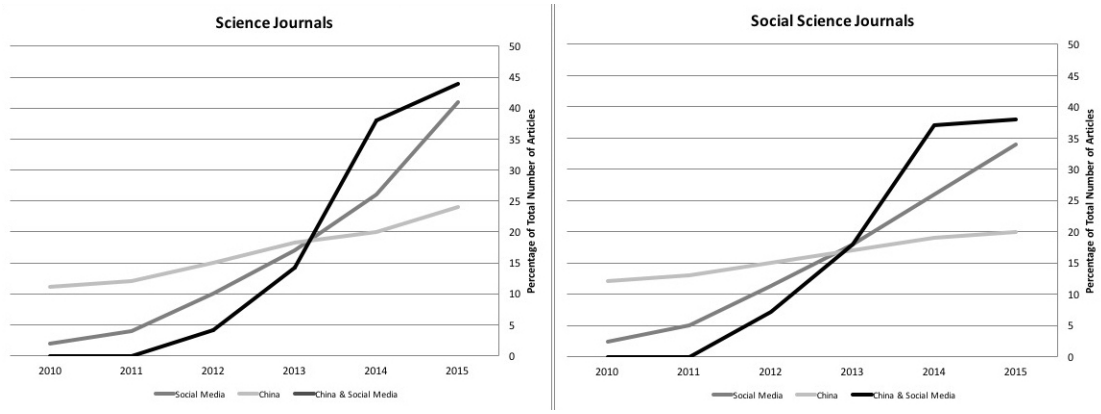
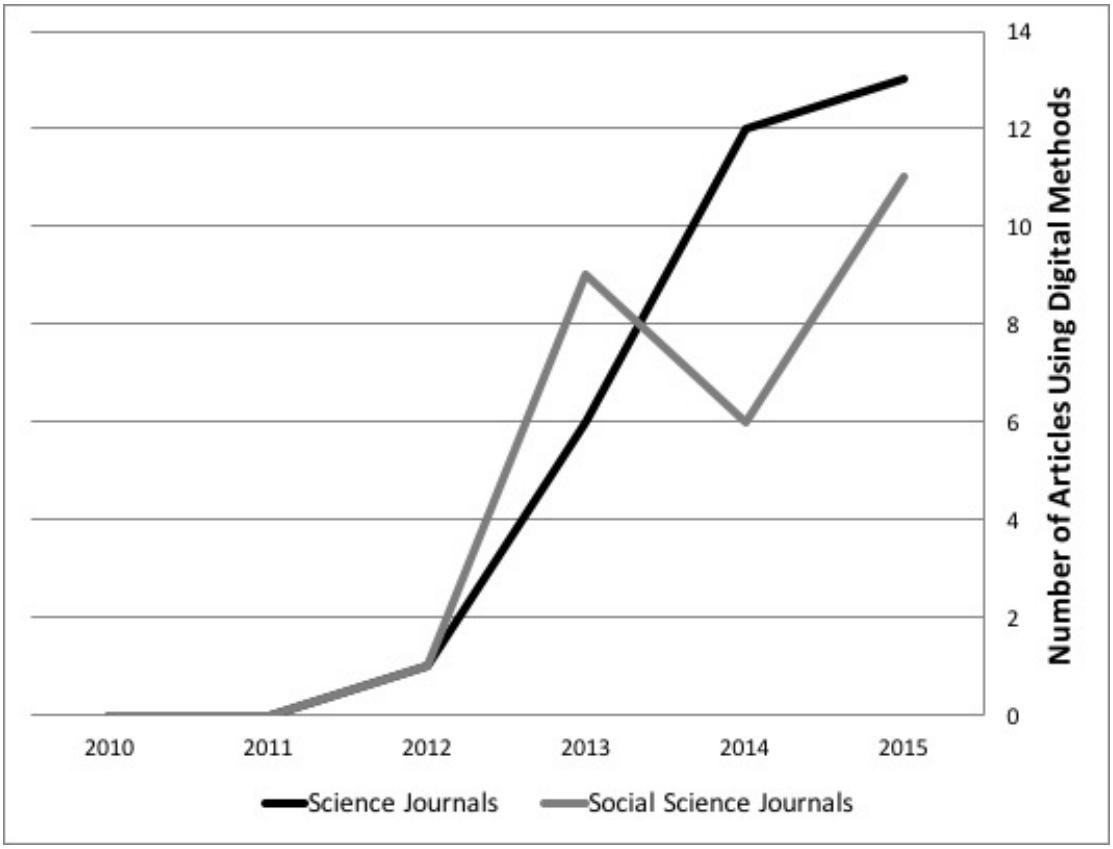


Figure 2. Number of Articles on Chinese Social Media Using Digital Methods for Data Collection or Analysis in Science and Social Science Journals. Source: Web of Science Citation Index, 2010-2015.



References

Abaza, M. 2011. "Academic tourists sight-seeing the Arab Spring." *Ahram Online* [online]. February 8, 2017. http://english.ahram.org.eg/News/22373.aspx.

Alsabah, N. 2016. "Nationale Sicherheit 2.0: Chinas Cyberspace-Behörde zähmt das Internet." *Merics China Monitor* merics.org: Mercator Institute for China Studies.

Argomaniz, J., O. Bures and C. Kaunert. 2015. "A Decade of EU Counter-Terrorism and Intelligence: A Critical Assessment." *Intelligence and National Security,* 30(2-3), 191-206.

Auer, M. and K.-w. Fu. 2015. "Clearing the air: investigating Weibo censorship in China: New research to show censorship of microbloggers who spoke out about pollution documentary." *Index on Censorship,* 44(3), 76-79.

Bamman, D., B. O'Connor and N. A. Smith. 2012. "Censorship and Deletion Practices in Chinese Social Media." *First Monday* [online], 17 (3). December 15, 2015. http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/3943/3169.

Bandurski, D. 2008. "China's Guerrilla War for the Web." *Far Eastern Economic Review* [online]. July 7, 2008. http://www.feer.com/.

Boas, T. 2006. "Weaving the Authoritarian Web: The Control of Internet Use in Nondemocratic Regimes." *In:* Zysman, J. and A. Newman eds. *How Revolutionary was the Digital Revolution: National Responses, Market Transitions, and LGobal Technology.* Palo Alto: Stanford University Press, 373-390.

Borgman, C. L. 2015. *Big Data, Little Data, No Data: Scholarship in the Networked World.* Boston: The MIT Press.

boyd, d. and K. Crawford. 2012. "Critical Questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon." *Information, Communication & Society,* 15(5), 662-679.

Brady, A.-M. 2008. *Marketing Dictatorship: Propaganda and Thought Work in Contemporary China.* Lanham: Rowman & Littlefield.

Cairns, C. and A. Carlson. 2016. "Real World Islands in a Social Media Sea: Nationalism and Censorship on Weibo (微博) during the 2012 Diaoyu/Senkaku Crisis." *China Quarterly,* 225, 23-49.

Chase, M. and J. Mulvenon. 2002. *You've got Dissent! Chinese Dissident use of the Internet and Beijing's Counter-Strategies.* Santa Monica, CA: Rand.

Chen, L., C. Zhang and C. Wilson. 2013. "Tweeting Under Pressure: Analyzing Trending Topics and Evolving Word Choice on Sina Weibo." *Proceedings of the 1st Annual ACM Confernce on Online Social Networks (COSN 2013).* Boston, MA.

Cheng, C. and M. Manion. 2015. "Beyond Censorship: Autocrats and Netizens in China." *Workshop on Multifaceted Study of Chinese Politics and Society.* School of International and Public Affairs, Shanghai Jiaotong University; Lieberthal-Rogel Center for Chinese Studies, University of Michigan.

Clark, W. R. and M. Golder. 2015. "Big Data, Causal Inference, and Formal Theory: Contradictory Trends in Political Science?: Introduction." *PS: Political Science & Politics,* 48(1), 65-70.

Clinton, A., Z. Lixuan and P. Iryna. 2014. "Investigating Privacy Perception and Behavior on Weibo." *Journal of Organizational and End User Computing (JOEUC),* 26(4), 43-56.

Conte, R.*, et al.* 2012. "Manifesto of computational social science." *The European Physical Journal Special Topics,* 214(1), 325-346.

de Weerdt, H. In Press. *Information, Territory, and Networks: The Crisis and Maintenance of Empire in Song China.* Cambridge: Harvard University Press.
Dodge, M. and R. Kitchin. 2005. "Codes of Life: Identification Codes and the Machine-Readable World." *Environment and Planning D: Society and Space,* 23(6), 851-881.
Drake, W. J. 1993. "Territoriality and Intangibility: Transborder Data Flows and National Sovereignty." *In:* Nordenstreng, K. and H. I. Schiller eds. *Beyond National Sovereignty: International Communications in the 1990s.* Norwood: Ablex, 259-313.
Fan, R.*, et al.* 2014. "Anger Is More Influential than Joy: Sentiment Correlation in Weibo." *PLoS ONE,* 9(10), e110184.
Ford, H. 2014. "Big Data and Small: Collaborations between ethnographers and data scientists." [online], 1 (2). February 8, 2017. http://bds.sagepub.com/spbds/1/2/2053951714544337.full.pdf.
Fu, K.-w., C.-h. Chan and M. Chau. 2013. "Assessing Censorship on Microblogs in China: Discriminatory Keyword Analysis and the Real-Name Registration Policy." *IEEE Internet Computing,* 17(3), 42-50.
Fu, K.-w. and M. Chau. 2013. "Reality Check for the Chinese Microblog Space: A Random Sampling Approach." *PLoS ONE,* 8(3), e58356.
Fu, K.-w. and M. Chau. 2014. "Use of Microblogs in Grassroots Movements in China: Exploring the Role of Online Networking in Agenda Setting." *Journal of Information Technology & Politics,* 11(3), 309-328.
Gallagher, M. E. and B. A. P. Miller. 2017. "Can the Chinese government really control the Internet? We found cracks in the Great Firewall." *Monkey Cage - The Washington Post*, p.
Garnaut, A. 2014. "The Geography of the Great Leap Famine." *Modern China,* 40(3), 315-348.
Gerlitz, C. and B. Rieder. 2013. "Mining One Percent of Twitter: Collections, Baselines, Sampling." *M/C Journal* [online], 16 (2). February 8, 2017. http://journal.media-culture.org.au/index.php/mcjournal/article/viewArticle/620.
Gerry Qc, F. and N. Berova. 2014. "The rule of law online: Treating data like the sale of goods: Lessons for the internet from OECD and CISG and sacking Google as the regulator." *Computer Law & Security Review,* 30(5), 465-481.
González-Bailón, S. 2013. "Social science in the era of big data." *Policy & Internet,* 5(2), 147-160.
González-Bailón, S.*, et al.* 2014. "Assessing the bias in samples of large online networks." *Social Networks,* 38, 16-27.
Greitens, S. C. 2013. "Authoritarianism Online: What Can We Learn from Internet Data in Nondemocracies?" *PS: Political Science & Politics,* 46(02), 262-270.
Gu, B. and Q. Ye. 2014. "First Step in Social Media: Measuring the Influence of Online Management Responses on Customer Satisfaction." *Production and Operations Management,* 23(4), 570-582.
Gu, H.*, et al.* 2014. "Importance of Internet Surveillance in Public Health Emergency Control and Prevention: Evidence From a Digital Epidemiologic Study During Avian Influenza A H7N9 Outbreaks." *Journal of Medical Internet Research* [online], 16 (1). http://www.ncbi.nlm.nih.gov/pubmed/24440770.
Han, R. 2015a. "Combating Corruption Online: Citizen Participation and State Responses in China " *Annual Meeting of the American Political Science Association.*
Han, R. 2015b. "Manufacturing Consent in Cyberspace: China's "Fifty-Cent Army"." *Journal of Chinese Current Affairs,* 44(2), 105-134.
Hintz, A. 2016. "Restricting digital sites of dissent: commercial social media and free expression." *Critical Discourse Studies,* 13(3), 325-340.

Jiang, M. 2010. "Authoritarian Informationalism: China's Approach to Internet Sovereignty " *SAIS Review of International Affairs,* 30(2), 71-89.
Jiang, M. 2013. "China's "Internet Sovereignty" in the Wake of WCIT-12." *China-US Focus* [online]. http://www.chinausfocus.com/peace-security/chinas-internet-sovereignty-in-the-wake-of-wcit-12/.
Katzenstein, P. J. 2001. "Area and Regional Studies in the United States." *PS: Political Science & Politics,* 34(04), 789-791.
King, G. 2014. "Restructuring the Social Sciences: Reflections from Harvard's Institute for Quantitative Social Science." *PS: Political Science & Politics,* 47(01), 165-172.
King, G., J. Pan and M. E. Roberts. 2013. "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review,* 107(02), 326-343.
King, G., J. Pan and M. E. Roberts. 2014. "Reverse-engineering censorship in China: Randomized experimentation and participant observation." *Science,* 345(6199).
King, G., J. Pan and M. E. Roberts. In Press. "How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, not Engaged Argument." *American Political Science Review*.
Kitchin, R. 2013. "Big data and human geography: Opportunities, challenges and risks." *Dialogues in Human Geography,* 3(3), 262-267.
Kozinets, R. V. 2010. *Netnography: Doing Ethnographic Research Online.* London: Sage Publisher.
Lazer, D.*, et al.* 2014. "The Parable of Google Flu: Traps in Big Data Analysis." *Science,* 343(6176), 1203-1205.
Lazer, D.*, et al.* 2009. "Life in the network: the coming age of computational social science." *Science (New York, N.Y.),* 323(5915), 721-723.
Li, G., B. Hao and T. Zhu. 2014. "How did the Suicide Act and Speak Differently Online? Behavioral and Linguistic Features of China's Suicide Microblog Users." *ArXiv e-prints, arXiv: arXiv:1407.0466* [online]. December 15, 2015.
Liang, H. and K.-w. Fu. 2015. "Testing Propositions Derived from Twitter Studies: Generalization and Replication in Computational Social Science." *PLoS ONE,* 10(8), e0134270.
Liu, Y.*, et al.* 2014. "Uncovering Patterns of Inter-Urban Trip and Spatial Interaction from Social Media Check-In Data." *PLoS ONE* [online], 9 (1).
MacKinnon, R. 2009. "China's Censorship 2.0: How Companies Censor Bloggers." *First Monday* [online], 14 (2). February 7, 2017. http://firstmonday.org/article/view/2378/2089.
MacKinnon, R. 2012. *Consent of the Networked: Tee Worldwide Struggle for Internet Freedom.* New York: Basic Books.
Marz, N. and J. Warren. 2012. *Big Data: Principles and Best Practices of Scalable Realtime Data Systems.* Westhampton: Manning.
Mayer-Schönberger, V. and K. Cukier. 2013. *Big Data: A Revolution that will Change How We Live, Work and Think.* London: John Murray.
Meissner, M. and J. Wübbeke. 2016. "IT-backed Authoritarianism: Information Technology Enhances Central Authority and Control Capacity under Xi Jinping." *In:* Heilmann, S. and M. Stepan eds. *China's Core Executive: Leadership styles, structures and processes under Xi Jinping.* 52-57.
Miller, B. A. P. 2016. "Automatic Detection of Comment Propaganda in Chinese Media." [online]. https://papers.ssrn.com/sol3/Papers.cfm?abstractid=2738325.
Monroe, B. L.*, et al.* 2015. "No! Formal Theory, Causal Inference, and Big Data Are Not Contradictory Trends in Political Science." *PS: Political Science & Politics,* 48(1), 71-74.

Morozov, E. 2011. *The Net Delusion: How Not to Liberate the World.* London: Penguin Books.
Morstatter, F.*, et al.* 2013. "Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose." *International Conference on Weblogs and Social Media (ICWSM).* 400-408.
Mostern, R. 2011. *Dividing the Realm in Order to Govern: The Spatial Organization of the Song State* Cambridge: Harvard University Press.
Mullins, J. 2007. "Enabling International Access to Scientific Data Sets: Creation of the Distributed Data Curation Center (D2C2)." *28th Annual Conference of the International Association of Technological University Libraries (IATUL).* Stockholm: Libraries Research Publications.
Nagler, J. and J. A. Tucker. 2015. "Drawing Inferences and Testing Theories with Big Data." *PS: Political Science & Politics,* 48(1), 84-88.
Nip, J. Y. M. and K.-w. Fu. 2016. "Challenging Official Propaganda? Public Opinion Leaders on Sina Weibo." *The China Quarterly,* 225, 122-144.
Niu, J.*, et al.* 2013. "An Empirical Study of a Chinese Online Social Network--Renren." *Computer,* 46(9), 78-84.
O'Brien, K. J. 2011. "Studying Chinese Politics in an Age of Specialization." *Journal of Contemporary China,* 20(71), 535-541.
Pan, J. 2017. "How Market Dynamics of Domestic and Foreign Social Media Firms Shape Strategies of Internet Censorship." *Problems of Post-Communism,* 64(3-4), 167-188.
Pentland, A. 2014. *Social Physics: How Good Ideas Spread - The Lessons from a New Science.* New York: The Penguin Press.
Pieke, F. 2013. *Contemporary China Studies in the Netherlands.* Leiden: Brill Publishers.
Qiu, J. L. 2000. "Virtual Censorship in China: Keeping the Gate Between the Cyberspaces." *International Journal of Communications Laws and Policy,* 4, 1-25.
Ran, B. 2013. "Use of Cellphone Data in Travel Survey and Transportation Planning." *Urban Transportation,* 11(1), 72-81.
Rogers, R. 2013. *Digital Methods.* Boston: MIT Press.
Schlaeger, J. 2013. *E-Government in China: Technology, power and local government reform.* London: Routledge.
Schlaeger, J. and M. Jiang. 2014. "Official microblogging and social management by local governments in China." *China Information,* 28(2), 189-213.
Schneider, F. 2015. "Searching for 'Digital Asia' in its Networks: Where the Spatial Turn Meets the Digital Turn." *AsiaScape: Digital Asia,* 2, 57-92.
Schroeder, R. 2014. *Big Data and the brave new world of social media research.*
Stockmann, D. and T. Luo. In Press. "Which Social Media Facilitate Online Public Opinion in China?" *Problems of Post-Communism*.
Stockmann, D. and L. Zheng. 2015. "Who is a PRC user? Comparing Chinese Social Media User Agreements."
Thompson Klein, J. 2004. "Prospects for transdisciplinarity." *Futures,* 36(4), 515-526.
Tromble, R. and D. Stockmann. 2017. "Lost Umbrellas: Bias and the Right to be Forgotten in Social Media Research." *In:* Zimmer, M. and K. Kinder-Kurlanda eds. *Internet Research Ethics for the Social Age: New Cases and Challenges.* New York: Peter Lang Publishers.
van Eeten, M. J. G. and M. Mueller. 2012. "Where is the governance in Internet governance?" *New Media & Society,* 15(5), 720-736.
Vis, F. 2013. "A Critical Reflection on Big Data: Considering APIs, Researchers, and tools as Data Makers." *First Monday* [online], 18 (10). February 7, 2017. http://firstmonday.org/ojs/index.php/fm/article/view/4878.

Wallach, H. 2016. "Computational Social Science: Toward a Collaborative Future." *In:* Alvarez, M. ed. *Computational Social Science: Discovery and Prediction.* New York: Cambridge University Press, 307-316.
Wang, L. 2015. *大数据领导干部读本 (Big Data Instructions for Cadres).* Beijing: Renmin Publishers (Renmin Chubanshe).
Wang, N., J. She and J. Chen. 2014. "How "Big Vs" Dominate Chinese Microblog: A Comparison of Verified and Unverified Users on Sina Weibo." *Proceedings of the 2014 ACM conference on Web science.* Bloomington, Indiana, USA: ACM, 182-186.
Watts, D. J. 2013. "Computational Social Science: Exciting Progress and Future Directions." *The Bridge on Frontiers of Engineering* [online], 43 (4). https://www.nae.edu/Publications/Bridge/106112/106118.aspx.
Welles, B. F. 2014. "On minorities and outliers: The case for making Big Data small." *Big Data & Society* [online], 1 (1). 2014-04-01 00:00:00. http://bds.sagepub.com/spbds/1/1/2053951714540613.full.pdf.
Welser, H.*, et al.* 2008. "Distilling Digital Traces: Computational Social Science Approaches to Studying the Internet." *In:* Fielding, N., R. M. Lee and G. Blank eds. *The SAGE Handbook of Online Research Methods.* London: Sage Publications, 116-141.
Wright, J. 2014. "Regional variation in Chinese internet filtering." *Information, Communication & Society,* 17(1), 121-141.
Wright, J., A. Darer and O. Farnan. 2015. "Detecting Internet Filtering from Geographic Time Series." *arXiv e-prints, arXiv: 1507.05819* [online]. 10 July 2017. http://arxiv.org/pdf/1507.05819v1.pdf.
Wu, L.*, et al.* 2014. "Intra-Urban Human Mobility and Activity Transition: Evidence from Social Media Check-In Data." *PLoS ONE,* 9(5), e97010.
Wu, Y.*, et al.* 2011. "A comparative study of online privacy regulations in the U.S. and China." *Telecommunications Policy,* 35(7), 603-616.
Yu, L. L., S. Asur and B. A. Huberman. 2012. "Artificial Inflation: The True Story of Trends in Sina Weibo." *ArXiv e-prints: arXiv:1202.0327v1* [online]. January 12, 2016. http://arxiv.org/abs/1202.0327.
Zeng, J. 2016. "China's date with big data: will it strengthen or threaten authoritarian rule?" *International Affairs,* 92(6), 1443-1462.