Article

new media & society
1–22

DOI: 10.1177/1461444820905068
journals.sagepub.com/home/nms
SAGE

Curbing profanity online: A network-based diffusion analysis of profane speech on Chinese social media

Yunya Song
Hong Kong Baptist University, Hong Kong

K Hazel Kwon
Arizona State University, USA

Jianliang Xu
Xin Huang
Hong Kong Baptist University, Hong Kong

Shiying Li
Brown University, USA

Abstract
Profanity, also known as swearing, refers to the use of foul language that is often linked to incivility. In Chinese digital space, the state government actively censors profanity under the rationale of protecting online civility. This study examines the diffusion of profanity in Sina Weibo, one of the largest Chinese social media platforms. The study applied computational methods to reconstruct the cascade networks of swearing and non-swearing posts and analyzed the network diffusion processes based on a set of structural metrics including reposting depth, width, and interlayer width ratios. Findings suggest profanity may influence the process of message diffusion, but this effect was ephemeral. Based on the understanding of diffusion processes of profanity online, this study contends the viral potential of profanity may not be as severe as the regulators

Corresponding author:
K Hazel Kwon, Walter Cronkite School of Journalism and Mass Communication, Arizona State University, 555 N. Central Avenue, Phoenix, AZ 85004, USA.
Email: khkwon@asu.edu

claim. The discussion analyzes the extent to which content moderation efforts are necessary for the nurturing of civility online.

Keywords
China, diffusion, online incivility, online propagation, profanity, social media, swearing

The use of foul language—also known as swearing or profanity—has become increasingly prevalent in digital spaces (Fägersten, 2017; Kwon and Gruzd, 2017). Although swearing may signal informality or relational closeness in some interpersonal communication contexts (Cavazza and Guidetti, 2014), social interactions in public networked platforms often occur among netizens who share few interpersonal relational histories (Arefi et al., 2019; Kwon and Gruzd, 2017). Swearing in digital public spaces is thus often associated with vulgarity and taken as an indicator of offensive and uncivil speech. While a healthy democracy relies on citizens' overall ability to form their opinions based on well-informed deliberation, uncivil speech has raised concerns about the decline of rationality in digital spaces (Anderson et al., 2016). In particular, unfiltered user-generated contents have become abundant in social media, thus the amount of profanity has also increased (Song and Wu, 2018). The rise of profanity in highly interactive social platforms has received increasing attention from scholars and practitioners alike, who strive to balance between protecting freedom of expression and curbing online incivility (Antoci et al., 2016).

The goal of this study is to understand the ways in which profanity propagates in China's social media environment. We explored China's popular social media service, Sina Weibo, by analyzing the reposting chains of Sina Weibo messages. As is widely known, China imposes hard-line digital governance and advocates for the state to have authority over the contents of its Internet domain. As a prominent proponent of "cyber sovereignty," the Chinese government subjects the Internet to the same intensity of regulation as Chinese life offline. The regulatory mechanisms have not been always effective, however, as the sheer volume of posts and netizens' creative use of code words to skirt censors have posed challenges for the regulators to identify and censor all such content (Song and Wu, 2018). In the context of growing tensions and controversy surrounding China's censorship of digital incivility, this study examines the virality of profane speech to understand to what extent profanity propagates differently from, or similarly to, "normal" messages in Sina Weibo. Empirically grounded understanding of the diffusion potential of profanity may help advance discussions about the needs and efficacy of the censorship of profanity in China.

Methodologically, this study used network-based diffusion analysis. Online diffusion processes involve variations in cascading structures, ranging from a large burst of one-step transmissions to a chain effect that involves multiple steps of concatenation. For example, a message may be broadcast to large number of users simultaneously, but the propagation may stop there. In this instance, the message quickly reaches a certain number of users but its virality is ephemeral. Conversely, the same number of users could be reached by the message's continuous flow through multiple steps, with each user passing it along to a subset of his or her networked friends (Goel et al., 2015). Therefore, although the viral speed might be slower than the one-time broadcasting, the message effect might endure longer and potentially generate storytelling variations.

Diffusion of social media content is the combined outcome of different modes of message flow. The current study analyzed message flow using the profane speech diffusion network data by reconstructing complete cascade networks of selected Sina Weibo posts. Also, this study disentangled swearing effects on the diffusion process from the effects of emotions and topics, which have been shown previously to influence online virality (Nahon and Hemsley, 2013; Vosoughi et al., 2018). Based on diffusion analysis results, this article discusses whether the viral potential of profanity is up to par in the way as claimed by the regulators and to what extent top-down content moderation maintains civility in Chinese digital life.

Background

Cyber governance in China: incivility and content controls

Online incivility has become a concern due to its potential troubling consequences. The conceptualization of incivility is twofold. On an interpersonal level, incivility refers to "violations of interpersonal politeness norms" (Muddiman, 2017: 3182). On a public level, incivility is defined as a violation of "democratic norms" (Muddiman, 2017: 3182) and is concerned with denying others the right to speak their minds, or the rejection of compromise or cooperation with people who hold different opinions (Muddiman and Stroud, 2017). In a digital space, incivility has manifested in various forms. For example, "flaming" pertains to disinhibited user speech with an intention to trigger inflammatory interactions (Cho and Kwon, 2015; O'Sullivan and Flanagin, 2003). Also, "cyberbullying" is a repeated hostile verbal attack with an intention to harass or threaten another person (Vandebosch and Van Cleemput, 2009). "Trolling" is another problem of incivility, which is an intent to disrupt social interactions, stir up disputes, or aggravate conflicts within a larger community using trickery (Coleman, 2014; Hardaker, 2010). Studies have found that online incivility has aggravated the polarization of political attitudes (Anderson and Huntington, 2017; Hwang et al., 2014) and disengagement with online discussions (Diakopoulos and Naaman, 2011).

Regarding the question of how to manage uncivil speech acts, implementing a moderation policy may encounter an even bigger ethical challenge because censure of incivility can operate as a silencing mechanism or create a means to subjugate subordinated groups (Jamieson et al., 2017). One notable concern is that a top-down implementation of politeness norms could be misused to justify excessive social control or free speech suppression (Jiang and Esarey, 2018). That is, governing actors can exploit the logic of civility as a rationale to pursue other social or political ends (Strachan and Wolf, 2012). The demand for civility may work "against a democratic order and in support of special interests, institutions of privilege, and structures of domination" (Kasson, 1990: 3) and "restrict content and participation through the limits [governing actors] place on acceptable style" (Ferree et al., 2002: 313–314). Policies that intend to promote civility can be at odds with the pursuit of democratic values, such as freedom of speech, that an invigorated liberalism necessitates (Kennedy, 1998).

China's digital governance is considered an exemplification of these ethical tensions (de Seta, 2018). China has grappled with the question of how to utilize Internet technology as a tool to advance its participation in global economy while simultaneously controlling byproducts of the techno-globalization, especially digital voices opposing the nation's political values (Wu, 2009). One context under which the Chinese authorities have imposed digital censorship has been that the concept that the Internet contains profanity—obscene and vulgar content—which is harmful to society and thus needs to be regulated (Wang, 2012). China's cyber-governance has been overreaching and proactive, characterized by a broad control over online materials, and emphasizes on building harmony while simultaneously policing dissenting opinions (Liang and Lu, 2012).

Studies of (in)civility in China have pointed to the nation's distinct socio-technical situation and political complexity (Yang, 2018). Chinese authorities publicly problematize the lack of civility in the nation's digital culture, using this as a justification for state control over the digital platforms. For example, the authorities have harnessed the so-called big data revolution, working closely with state-licensed commercial tech giants (e.g. Alibaba, Baidu and Tencent) to build systems of algorithmic surveillance. Official agencies are responsible for monitoring online content, and Internet content providers are required to implement self-censorship mechanisms (Hou, 2018). However, the authorities allow incivility to go unchecked at times. For example, some polarized and offensive comments touting nationalistic views are not deleted as they are considered to be in accordance with the objective of curbing "penetration of Western values" and preserve "cultural security" (de Seta, 2018: 2018). While uncivil speech—varying from online vigilantism, trolling, and verbal profanity—have been regulated by the authorities under the rationale of civility, a porous conceptualization of civility has made the boundary of censorship ambiguous.

Studies have shown that China's civility campaigns and subsequent censorships have often framed an expression of high emotion such as anger as irrational and uncivil, thereby weakening online mobilization and contentious collective actions through Chinese digital networks (King et al., 2013; Yang, 2018). Government censorship screens the cyberspace, deleting words and phrases dubbed inflammatory or seditious. To counteract, Chinese netizens use profanity as a means of subversion (Wang, 2012). For example, Chinese netizens spread a video meme that illustrated a grass-mud horse defeating a river crab. Here, the word "grass-mud horse" is pronounced in the same manner as a popular Chinese profanity "fuck your mother," and the word "river crab" was the homonym to the government's propaganda catchword "harmony," representing the official line that calls to silence dissenting voices. By using nuanced profanity, the video meme mocked the clean-up campaigns, exemplifying netizens' discontent with Internet policing and censorship (Tang and Yang, 2011; Wang, 2012).

The tension between government's logic of civility underlying the Internet control and netizens' resistance against censorship practices on social networks is contentious in China's digital politics. Since the "Special Campaign to Rectify Vulgar Content on the Internet" campaign in early 2009, curbing profane and vulgar content has been a major mission in China's regulatory endeavors to secure digital civility (Jiang and Esarey, 2018). However, the regulator's insinuation that the spread of profanity is entrenched in the digital space has *not* been attested by empirical evidence. In this study, we empirically examined profane speeches on Sina Weibo as examples of "vulgar content" per the regulator's definition of incivility. Specifically, we conducted a network-based diffusion analysis of profane speech, in comparison with that of "normal" content that did not pose a concern for incivility. The aim was to address whether the virality of profanity is farther and wider than that of ordinary posts and whether its propagation in the digital space is to such an extent that it is dangerously viral to disrupt social order.

The spread of profane speech online

This study examines diffusion of profane speech in Sina Weibo. Although profanity has been related to the concept of incivility, this study does not claim that profanity equates to incivility. Rather, profanity is one of the most commonly employed linguistic tactics used to trigger emotional arousal and a sense of offensiveness (Jay, 2000; Kwon and Cho, 2017; Mak and Lee, 2015). While prior empirical studies often examine profanity as one type of uncivil speech (e.g. Anderson et al., 2016; Kwon and Gruzd, 2017), some scholars have pointed out that profanity is either a minor form of incivility or even not considered a distinctive marker of incivility in some contexts (Jay and Janschewitz, 2008; Kenski et al., 2017; Stryker et al., 2016). A politeness-focused definition does not fully capture all aspects of civil public discussions in that heated discussions often turn out to be emotional or impolite. That said, incivility has often been operationalized as the level of profanity and the presence or absence of aggressiveness (Boatright et al., 2019). Moreover, the perception of incivility is sensitive to contextual factors. In the Chinese context, curbing profanity has been at the forefront of the state-managed Internet clean-up campaign for digital civility (Yang, 2018).

Previous studies on online message diffusion have paid attention to message characteristics and explored how such characteristics influence the intensity of message transmission. For example, emotions were found to be one of the strongest predictors of online message transmission. Results regarding which emotional valence has greater viral potential have been mixed, with some studies suggesting a larger effect of positive emotion on online transmission rates than negative emotion (Berger and Milkman, 2012) and whereas other studies conversely identified a stronger effect of negative emotion (Ferrara and Yang, 2015; Song et al., 2016; Song and Xu, 2019; Vosoughi et al., 2018). Regardless, studies commonly agreed that a high emotional arousal makes a message more viral than emotionally neutral messages (Stieglitz and Dang-Xuan, 2013), alluding that the linguistic device of swearing that expresses a high level of emotion should influence the likelihood of message transmission.

Another well-studied message characteristic is the message topic. Kwon and Cho (2017) showed that the swearing effect on mobilizing online public's attention was moderated by news topics, in that the effect was more prominent for political topics than non-political topics. Vosoughi et al. (2018) examined a different incivility problem—fake news—and found that the diffusion of fake news was more likely "farther, faster, deeper, and more broad" when the news story contained a political issue (p. 1146).

Other research has examined the propagation of online profanity as a social process. Studies have found that online users enhance a sense of group affiliation or achieve other social relational goals by imitating other users' speech acts, including by swearing (Gonzales et al., 2010; Jefferis et al., 2003; Taylor and Thomas, 2008). In a similar vein, Kwon and Gruzd (2017) found the evidence of swearing contagion in the context of political discussions on YouTube, finding the spillover effect of swearing from a preceding comment to its following comments. Kim and Herring (2018) also found the contagious potential of politeness strategy in online commenting communities. Polite comments tended to be succeeded by polite comments, and impolite comments succeeded by impolite ones. Furthermore, the (im)politeness effect interacted with other social interactional rules such as anonymity (Kwon and Cho, 2017; Suler, 2004) and gendered norms (Kim and Herring, 2018).

While prior research enriches the understanding of content and social aspects of the profanity propagation, an analysis of diffusion network structure may offer further evidence about the community-wide impact of profane speech. Network-based diffusion analysis helps track how a message is propagated within the whole community (Liu-Thompkins, 2012). Furthermore, an analysis of the full cascade network data allows measurement not only of the sheer diffusion size (i.e. total number of reposts) but also more refined structural features such as depth (i.e. the penetrability of a message), width (i.e. the expansibility of a message), and interlayer width rate (i.e. the expansion rate of the message propagation) (Liu et al., 2017). Different measures of diffusion rates provide understanding of a message's diffusion pattern which may fall somewhere in a continuum between a complete broadcast (i.e. a message reaches the audience through a single one-to-many broadcast) and a complete word-of-mouth (i.e. a message spread by multiple generations with each participating individual responsible for a small fraction of the total reach) (Goel et al., 2016).

Diffusion of profanity has not been studied from a network approach probably due to the restricted access to complete network data, especially in the context of China's restrictive social media environment. To study the virality of online profane speech, the current study followed a similar approach taken by a recent fake news diffusion study (Vosoughi et al., 2018), which compared the diffusion of true and false news on Twitter by constructing cascade networks of 126,000 news items. This study (2018) analyzed news diffusion patterns based on the various network structural features, concluding that "the spread of falsehood was aided by its virality, meaning that falsehood did not simply spread through broadcast dynamics but rather through peer-to-peer diffusion" (p. 1148). Likewise, the current study compared the diffusion processes between swearing and non-swearing posts on Sina Weibo.

Research question and designs

This study poses the question: *How does the diffusion process of profane speech differ from the diffusion process of non-profane speech?* To address this question, we analyzed a large corpus of swearing Weibo posts (N=1,222,862) by 875,075 users collected from 21 June to 18 July, 2017. The data collected in this study are all publicly available on the Sina Weibo site.

Sina Weibo is one of the most popular social networking sites in China. Since its launching in 2009, Sina Weibo had attracted over 486 million monthly active users by June 2019, which accounted for 56.9% of the total Internet users in China (CNNIC, 2019). The functionality of Weibo is similar to Twitter, such as reposting, mentions, hashtags, and URL shortening. Users provide detailed information (e.g. gender and geographic location) in their publicly available profiles, follow other accounts, and post or repost short messages to their networks of followers. Although the post is controlled in 140 characters for non-member users, 140 Chinese characters can contain more information than English of the same length. Furthermore, Weibo allows posts with photos and videos and adopts a mechanism of threaded comments, which Twitter had not used until 2013. The threaded comments design is considered to add a dimension of interactivity and engagement, with 80.3% of its users following current news and social events on the platform (Song and Xu, 2019).

In this study, all publicly available search results from the Web interface of Weibo's search engine were retrospectively collected by searching for posts containing at least one of 62 Chinese profane words frequently used online. The list of profane words was created by compiling a list of the top vulgar words on Sina Weibo according to People's Daily Online Public Opinion Monitoring Center, and a list of vulgar words the news media and online platforms should avoid using released by the official Xinhua News Agency. For example, the following post contained swearwords and thus was included in our data corpus, with a reposting count of 14,809 times:

> Let me carefully explain how stupid the article written by the so-and-so League is. The metaphors in Dahufa [a Chinese animated film] are explicit, with fasces, Nazi salute, and the Rising Sun Flag alluding to Fascism. It turned out that the League put this label on itself . . . Indeed, there're plots such as prohibition of speech, but you can't just TMD (fucking) admit this . . .

Note that some contents are automatically filtered by the Sina Weibo server immediately upon submission, prohibiting the message from even being published in the first place, and thus it cannot be diffused.

Ideally, the full reposting networks would be constructed for all original posts in the data corpus. However, retrieving a full cascade of "who reposted whom" is computationally burdensome. As such, we systematically selected a small subset of original swearing posts from the collected corpus and constructed their full cascade networks. For comparison purposes, we also collected an equal number of non-swearing posts with similar contents and similar repost frequency distributions and constructed their cascade networks. Specific procedures are described below.

Emotion-based stratified sampling

As for the sampling strategy, we first took a stratified random sampling method based on the posts' emotional valence. Prior studies have found the important role of emotion in the virality of online content (e.g. Berger and Milkman, 2012; Stieglitz and Dang-Xuan, 2013), suggesting the need to control for the effect of emotional valence. That said, a purely random sampling would leave unknown how many positive or negative emotional messages were included in our dataset. It was of particular concern in this study because swearing is inherently an emotion-invoking utterance, especially often associated with a negative emotional state such as anger. As such, we first classified the collected swearing posts according to its emotional valence before taking additional sampling steps. The stratified sampling from emotion classes was intended to ensure a reasonable sample size for each emotional valence.

Emotional classification was performed by using supervised machine learning since the data corpus was too large to be handled manually. A convolutional neural network (CNN) algorithm was employed for machine learning process. Neural network–based algorithms, such as CNN or recurrent neural networks (RNN), are one of the most notable developments in machine learning techniques for text and image classification. With a large set of well-labeled training data, CNN results in highly accurate predictions. Compared to traditional machine learning methods such as support vector machine (SVM), CNN is known for better performance in sentence classification (Zhang and Wallace, 2015).

We trained an emotion classifier that automatically classified each post into one of four emotional categories including happiness, fear, anger, and sadness. Happiness was a positive emotion, while the remainder were negative emotions. After data cleaning and tokenization, we fed Weibo posts into the Word2vec model, which allows to capture syntactic and semantic relationships between words, as well as the context of words in a document. Specifically, the model generated a vector space from the text corpus, with each unique word being assigned a corresponding vector in the space. The length of a vector equaled a count for each word in the entire corpus. After feeding the one-hot representation[1] of that word into the network, we obtained the word embeddings for the target words by extracting the hidden layer that encoded the word representation. The output of the hidden layer was the word vector for the input word whose length was equal to 300.

Supervised machine learning requires training and testing datasets. The CNN classifier was trained using the labeled Weibo posts provided by the Natural Language Processing and Chinese Computing Association (NLPCC). Before feeding the training data into the algorithm, each word in each training Sina Weibo post was mapped to its corresponding word embeddings. The mapped input posts then passed through convolution layers, pooling layers, and fully connected layers that extracted, reduced, and combined semantic features to eventually classify the output into one of the four emotions. The convolution layer is the main building block of the CNN, and its function is to extract features from the original output. The function of pooling layers is to reduce the amount of parameters and computation, and fully connected layers multiply the input by a weight matrix and generate the output. The data collection and processing was carried out using Python 2.7. The python libraries used include "Jieba," "TensorFlow," "Gensim," and "sklearn."

The testing dataset was constructed using the labeled swearing post data. The model tests result that the categories of happiness and anger showed high precision (.76 for happiness and .73 for anger), recall (.82 for happiness and .86 for anger), and F1-score (.79 for both happiness and anger), while fear and sadness showed much lower precision (.31 for fear and .19 for sadness), recall (.11 for fear and .58 for sadness), and F1-score (.27 for fear and .37 for sadness). Furthermore, after classification of the whole corpus, most posts fell into either the category of happiness (49.51%, $N=605,398$) or anger (43.22%, $N=528,486$). Only 2.31% and 4.96% of posts were classified as either fearful or sad. Notably, the classification results showed that swearing occurred in positive emotional posts as frequently as in negative posts. Based on the results, we referred to the posts classified as happiness and anger to exemplify positive and negative emotions, respectively, and moreover these two emotion categories were used as controls in the analysis below.

Sampling original posts of varying network size

Within each emotional category, we pooled the sample posts representative of varying diffusion network size. For this task, the list of original swearing posts was rank-ordered by their repost frequencies. Finally, we chose 1% of posts from each rank-ordered list by selecting every 90th post from the list of negative posts, and every 110th post from the list of positive posts. This process resulted in a total of 20 negative and 60 positive-swearing posts with varied repost frequencies (minimum = 96; maximum = 16,830). We also manually validated that the selected swearing posts were indeed read as profane speech.

After selecting the original swearing posts, we manually collected a comparable set of non-swearing posts. For each negative-swearing post, we identified a post with a similarly negative emotional tone and a similar repost size, but one that did not show a sign of profanity, thereby creating a set of "negative-non-swearing posts." The same procedure was performed to create a set of "positive-non-swearing posts." As such, a total of 160 posts (20 negative-swearing, 20 negative-non-swearing, 60 positive-swearing, and 60 positive-non-swearing) were identified, and their diffusion networks were reconstructed.

Topic

In addition to emotion, topical effect was considered. Using manual coding, a binary variable was created as to whether the post contained political topics (=1) or not. Specifically, if a Weibo post mentions (1) officials, political appointees, or government employees; or (2) government agencies or departments at any level; or (3) government policies, decisions, or actions; or (4) issues in local, national, or international politics (e.g. diplomacy/defense/security issues), it was coded as a political post. Otherwise, it was coded as a non-political post. The inter-coder reliability was Cohen's kappa .98.

Diffusion network and measures

An original post and the subsequent chains of reposts comprise a diffusion tree network (Goel et al., 2016). For each original post, a diffusion network was constructed based on the poster's information and their reposting relationships. In each diffusion network, users (posters) were represented as nodes. The "root node" was the source user who posted the original message, and "leaf nodes" were the remaining users who reposted the message. The relationship between a pair of users was defined by their reposting relationship. A directional link E_{ij} was created if user i reposted user j's post. For each original post, a complete diffusion tree network was generated by starting

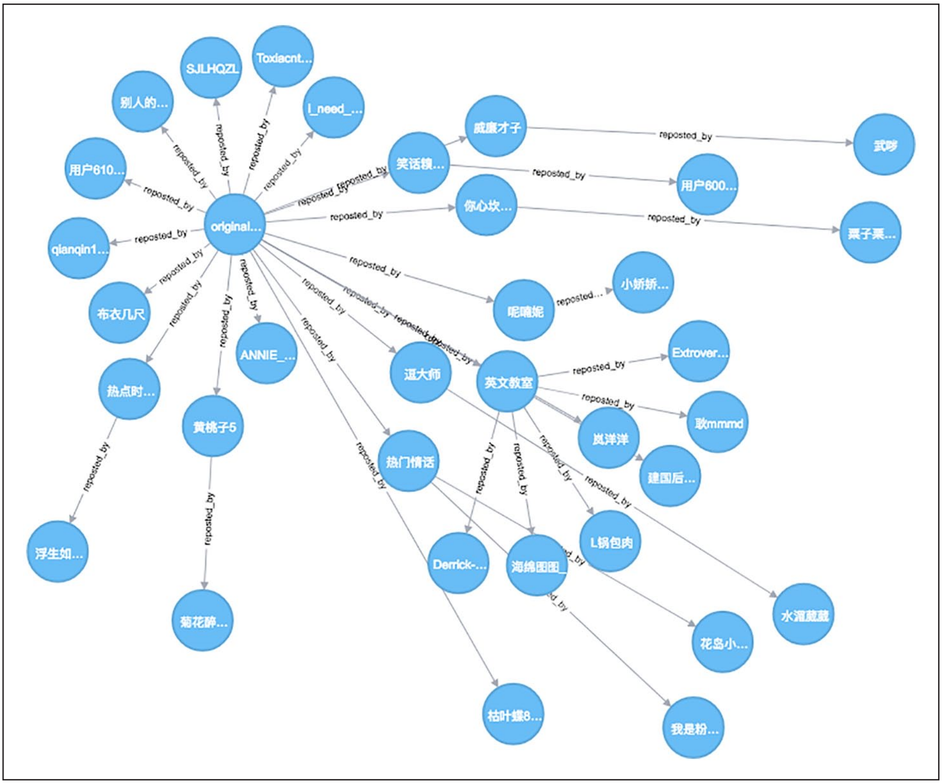


Figure 1. An example of an original Sina Weibo post's diffusion network.

from the original author and concatenating reposters as successive nodes. A total of 120 diffusion networks were constructed. Figure 1 visualizes an example of diffusion tree networks.

Online diffusion may be a product of different modes of message transmissions. To understand the ways in which an original post was propagated, we measured four properties of the diffusion tree networks including size, width, depth, and interlayer width ratios. Specifically, (1) the *size* of a network referred to the total number of reposts of the original message; (2) the *depth* of a network was the longest path from the root node to leaf nodes; (3) the *width* of a network was defined as a number of nodes in the layer that includes the largest number of nodes; and (4) an *interlayer width ratio* quantified the expansion rate of the message propagation, computed as the sum of the ratios between the number of nodes of two consecutive "layers," that is, the generations of reposters. A larger score of interlayer width ratio indicates a faster expansion rate of a message. The interlayer width ratio was computed up to the fifth generation. This study distinguished two types of interlayer width ratios. First, *interlayer width ratio 1* referred to the sum of the ratios between the number of nodes of two consecutive layers lower than the fifth generation (Liu et al., 2017):

$$L1(T) = \sum_{k=0}^{4} \frac{l_{k+1}}{l_k}, \tag{1}$$

where $L1(T)$ refers to the interlayer width ratio 1 of a diffusion network T, and l_k means the kth layer. Second, *interlayer width ratio 2* refers to the sum of the ratios between the number of nodes of two consecutive *non-root* layers lower than fifth generation:

$$L2(T) = \sum_{k=1}^{4} \frac{l_{k+1}}{l_k}, \tag{2}$$

where $L2(T)$ refers to the interlayer width ratio 2 of a diffusion network T. The interlayer width ratio 2 is understood as an expansion rate that comes after the most immediate transmission of an original message (i.e. from a root node to its first generation). The analogy of the two-step flow of communication model (Katz, 1957) may be used to conceptualize the interlayer width ratios: the interlayer width ratio 1 is analogous to a message's expansion rate across all steps including media broadcast (step 1), and word-of-mouth transmission from opinion leaders to followers and between followers (step 2 and further). In comparison, the interlayer width ratio 2 excludes the immediate media effect and pertains only to the message's expansion rate during the word-of-mouth transmission stage.

Analysis plan

First, a series of two-sample Kolmogorov–Smirnov (K-S) tests were performed to examine whether the distribution of each diffusion metric—size, width, depth, and interlayer width ratios—was different between the networks of swearing posts and those of non-swearing posts. Second, to further examine the effect of profanity on each diffusion metric, regression analysis was performed, adding emotion and topic as control variables. Considering that the diffusion metric variables were skewed, they were then log transformed. Stata 13 was used for the K-S tests and regression modeling.

Results

Descriptive statistics

The mean network size of the swearing and non-swearing posts was 1776.36 and 2136.14, respectively. The mean depth for swearing was 7.23 and non-swearing posts 8.58. The mean width for swearing was 1239.87 and non-swearing posts 1477.25. The mean interlayer width ratio 1 for swearing was 19.14 and non-swearing posts 1.42, while the mean interlayer width ratio 2 was .54 and .55, respectively.

The pairwise correlations (based on log-transformed values) showed that the network size and width were very highly correlated ($r = .98$, $p < .001$). The network size was also positively associated with depth ($r = .60$, $p < .001$), interlayer ratio 2 ($r = .18$, $p < .05$), and positive emotion ($r = .19$, $p < .05$). Given that the network size is the most basic

Table 1. Correlations among variables.

	1.	2.	3.	4.	5.	6.	7.
1. Topic	–						
2. Swear	–.08	–					
3. Emotion	.63***		–				
4. Size	.09	–.18*	.19*	–			
5. Depth	.27***	–.13	.42***	.60***	–		
6. Width	.03	–.18*	.10	.98***	.53***	–	
7. Interlayer 1	.15	.38***	.25**	.05	.14	–.01	–
8. Interlayer 2	.20*	–.04	.31***	.18*	.48***	.10	.36***

*$p < .05$; **$p < .01$; ***$p < .001$.

metric informing the degree of a diffusion network, we controlled the network size when running regression analyses for swearing effect on other metrics. Table 1 shows the correlations among variables.

K-S tests

The K-S tests examined the equality of distribution for each diffusion variable between swearing and non-swearing networks. Each K-S test answered two hypotheses: whether swearing networks contained a larger value of a given diffusion metric than non-swearing networks (D_1); and conversely, whether non-swearing networks contained a larger value than swearing networks (D_2).

The K-S results suggested that non-swearing networks showed both a larger network size ($D_2 = -.23$, $p < .05$) and larger width ($D_2 = -.24$, $p < .05$), while depth was not significantly different between the two groups. Interlayer width ratio 1 was the only metric in which swearing networks showed a larger value than non-swearing networks ($D_1 = .33$, $p < .001$). The descriptive statistics and the results of K-S tests are presented in Table 2. Also, Figure 2 visually compares the distribution of each diffusion metric between swearing and non-swearing networks.

Regression analysis

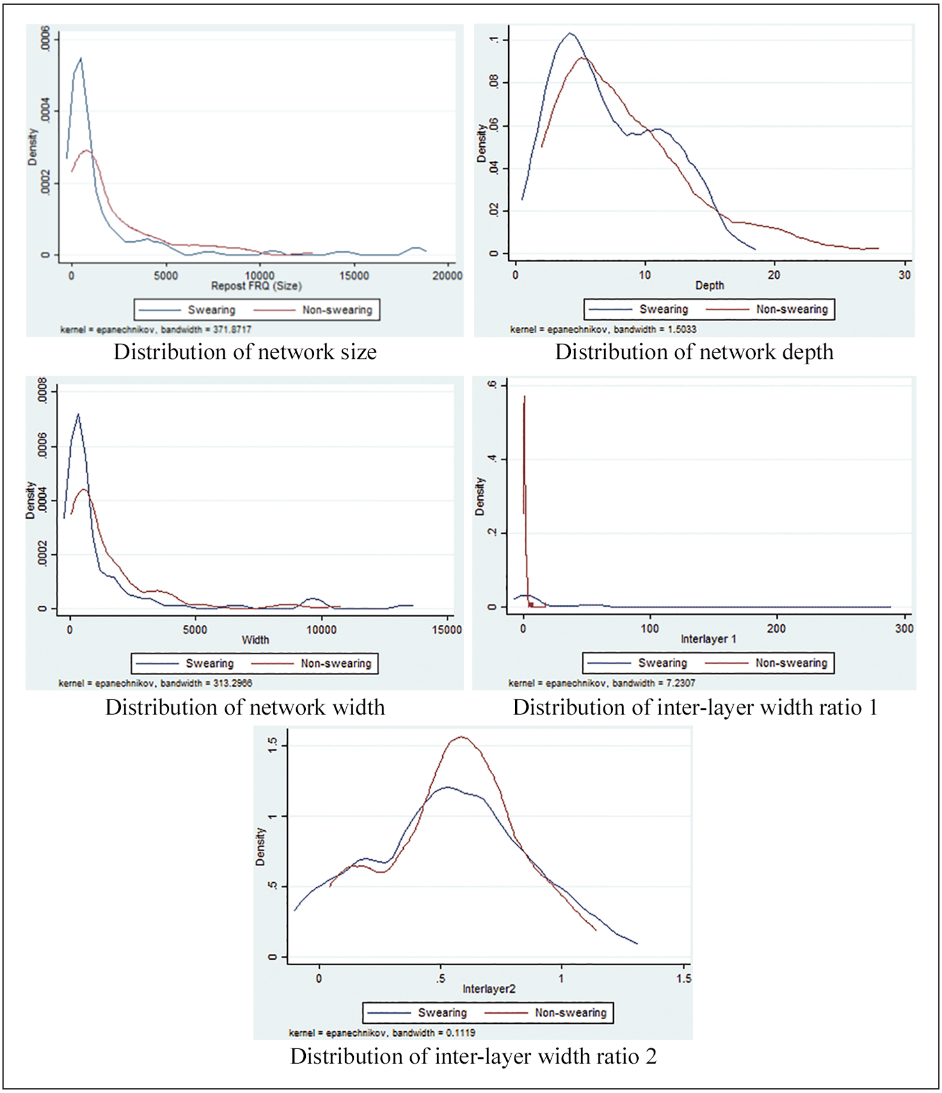
The ordinary least squares (OLS) regression models were performed to predict the effect of swearing on each diffusion metric after controlling the effect of emotion (anger = 1; happiness = 0), topic (politics = 1; non-politics = 0), and diffusion network size. Table 3 summarizes the results.

Swearing had a significant negative effect on the diffusion size, $b = -0.45$, $t(156) = -2.09$, $p < .05$. If a swearword occurred in an original post, the number of reposts was likely to be smaller. Next, controlling the diffusion size, we tested the effect of swearing on other diffusion variables. With respect to depth, swearing had an insignificant negligible effect, $b = -0.01$, $t(155) = -0.19$, $p = .85$. Likewise, swearing had an insignificant effect on width, $b = -0.01$, $t(155) = -0.12$, $p = .91$.

Table 2. Descriptive statistics and Kolmogorov–Smirnov (K-S) tests.

Variable	Non-swearing				Swearing				K-S test	
	M	*SD*	Min	Max	*M*	*SD*	Min	Max	D_1	D_2
Size	2136.14	2592.84	16.00	12,815.00	1776.36	3507.31	85.00	18,494.00	.04	–.23*
Depth	8.58	5.38	2.00	28.00	7.23	4.01	2.00	17.00	.00	–.11
Width	1477.25	1987.96	15.00	10,793.00	1239.87	2435.28	52.59	13,372.31	.04	–.24*
Interlayer 1	1.42	2.19	0.04	17.83	19.14	39.15	0.01	282.54	.33***	–.06
Interlayer 2	0.55	0.27	0.04	1.14	0.53	0.32	0.01	1.20	.06	–.11

D_1 tested a hypothesis that swearing networks contained a larger value than non-swearing networks; D_2 tested a hypothesis that non-swearing networks contained a larger value than swearing networks.
*$p < .05$; ***$p < .001$.



Distribution of network size

Distribution of network depth

Distribution of network width

Distribution of inter-layer width ratio 1

Distribution of inter-layer width ratio 2

Figure 2. Kernel density plots for distribution of diffusion metrics.

The only significant effect of swearing was found with respect to interlayer width ratio 1, $b=0.99$, $t(160)=5.4$, $p<.001$, consistent with the K-S test above. That is, swearing posts showed a fast expansion rate as soon as they departed from the root node. The comparison of swearing and non-swearing posts in terms of interlayer width ratio 1 showed a clear difference, as seen in Figure 3. Overall, the swearing networks had a higher interlayer width ratio 1 than the non-swearing networks across a varying spectrum of network sizes (i.e. reposting frequencies). This result suggests that profanity could prompt users of the first generation, who had immediate exposure to the original post, to

Table 3. Regression analysis for predicting diffusion of Sina Weibo posts ($N = 160$).

	Size		Depth		Width		Interlayer 1		Interlayer 2	
	b	*t*	*b*	*t*	*b*	*t*	*b*	*t*	*b*	*t*
Swearing (= 1)	–0.45*	–2.09	–0.01	–0.19	–0.01	–0.12	0.99***	5.4	0.00	–0.01
Topic (politics = 1)	–0.29	–0.67	0.09	0.68	–0.03	–0.35	0.16	0.44	0.02	0.27
Emotion (anger = 1)	0.75*	2.32	0.32***	3.41	–0.26***	–3.96	0.61*	2.25	0.12*	2.55
Size			0.21***	8.77	0.99***	59.14	0.08	1.14	0.02	1.49
Constant	6.77	41.11	0.60***	3.61	–0.20	–1.73	0.02	0.04	0.26**	3.22
R^2	0.06		0.45		0.96		0.22		0.10	
Adjusted R^2	0.04		0.44		0.96		0.20		0.08	

b = unstandardized coefficients.
$^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

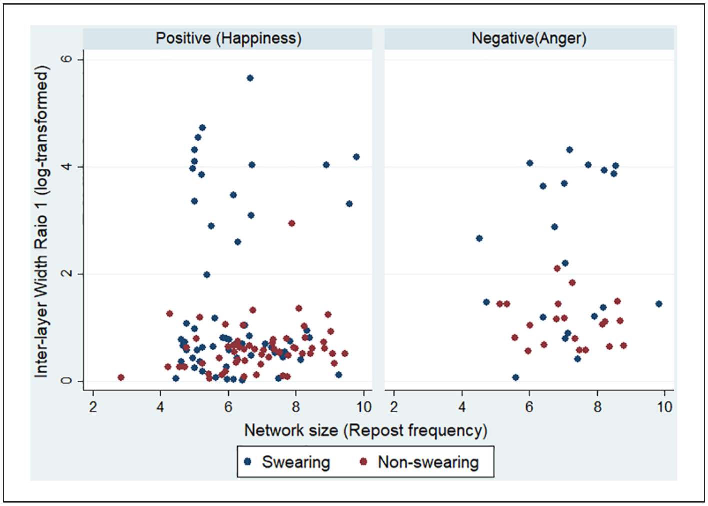


Figure 3. Effect of swearing on interlayer width ratio 1 by different emotion categories.

spread it. However, with the interlayer width ratio 2 as a dependent variable, swearing did *not* show any effect, $b=-0.00$, $t(160)=-.01$, $p=.995$, meaning that the over-time propagation of swearing posts after the first generation was not different from that of non-swearing posts. Together, interlayer width ratios suggest swearing posts might disseminate quickly at the initial stage, but the effect was ephemeral.

Among control variables, emotion was a significant predictor of all diffusion variables. Specifically, posts containing anger engendered more reposts (i.e. the size of the diffusion network) and went deeper (in terms of depth) with larger expansion rates (in terms of interlayer ratio 1 and 2) than those containing happiness. Happiness was more effective than anger only for increasing width. Meanwhile, political topic was not associated with any of the diffusion metrics.

Discussion and conclusion

This study examined the effect of profanity on the diffusion of online messages on China's popular microblogging site, Sina Weibo. While most of prior studies on online profanity focused on the US or European contexts, only a few studies exist in an Asian context (e.g. Cho et al., 2012; Cho and Kwon, 2015; Kwon and Cho, 2017). This study sought to fill this gap by analyzing online profane speeches in a Chinese digital space where the government implements top-down content moderation polices under the rationale of protecting online civility. By understanding the ways in which profane speeches are propagated on a popular Chinese social site, this study intends to provide data-driven, empirical grounds for further discussions about implications of top-down approaches to content moderation.

This study analyzed the propagation networks of Sina Weibo posts using diffusion analysis. A variety of network structural metrics were introduced to quantify multi-faceted virality including diffusion size, width, depth, and interlayer width ratios. While the diffusion network size captured the overall popularity of a post based on the total frequency of reposting, other metrics shed light on the process by which a post goes viral. Specifically, depth reflected the word-of-mouth process of diffusion, implying the longevity of message transmission chains. Meanwhile, width measured a message's broadcasting potential, that is, the message became widespread by involving a large population of adopters simultaneously. Furthermore, the interlayer width ratios measured a rate of change in virality between the prior generation of adopters and the successive generation. By examining various diffusion metrics, this study quantified the multiple aspects of online propagation. This enabled us to assess not only the volume of virality but also the generative processes of diffusion.

A central question of this study was to examine whether profane speech diffuses in a distinctive manner from ordinary messages such that it in fact reflects a disruptive potential and is thus worth controlling. Ostensibly, the results suggest that the swearing posts indeed showed a faster propagation than non-swearing posts. The result of interlayer width ratio 1 indicated that the propagation rate of swearing posts was significantly higher during the first-generative cycle (i.e. from the root node to its immediate neighbor nodes) than non-swearing posts. Prior studies of YouTube videos (e.g. Yoganarasimhan, 2011) suggest that both first- and second-degree nodes are equally important in deciding the virality of online content. This study found that the spread of profanity in Weibo showed a similar pattern in that the first generation is of unparalleled importance for the initial take off and to spread the word early on. Because the propagation rate for profane posts is relatively higher in early generations, it is advisable to devote resources to nipping it in the bud from a managerial or policy perspective. Conversely, given that non-profane posts tend to persist through a longer chain of reposting, netizens who intend to make impactful voices should be strategic in curating eye-catching content while keeping their conversations civil.

However, setting the first-generative cycle aside, profanity did *not* predict changes in any other diffusion metrics, implying that the first-generative propagation did *not* have an enduring effect. In fact, the K-S test results even suggested that swearing posts were associated with *smaller* total diffusion size and width.

Based on our findings, we conclude that the diffusion effect of profanity is ephemeral at best, and thus the overemphasis on curbing vulgarity as a reason for content control may be only valid at face value. The manifold, contested notion of profanity used by the regulators has alluded to the reasoning of civility in a rather ambiguous manner. For example, China's periodic campaigns to fight against "vulgar" online culture have been based on a loosely defined set of terminologies such as obscenity, vulgarity, and harmfulness by roughly referring to them as content harmful for minors (Li, 2011). With the exception of child pornography, other types of vulgar content have not been concretely specified. Such ambiguity would risk disciplining netizens with self-censorship. The current findings of ephemeral virality of profane speech lead to a cautious skepticism about the efficacy of profanity censorship and deletion practices on Chinese social media platforms.

While previous literature on swearing and profane speech has mostly taken a social psychological approach (e.g. treating swearing as emotional stimuli), this study intends to expand societal and policy discussions surrounding profanity, incivility, and online content moderation. In this sense, the network-based diffusion analysis helps broaden the analytic scope of this line of research.

That said, this study is not free from limitations. Most importantly, the current study centers around diffusion analyses that measure whether a profane post spreads in a different pattern from a non-profane post. While diffusion analysis helps uncover the propagation capability of posts, it does not attest to socio-cultural implications of profanity. As previous research has discussed, digital incivility could problematically influence minds and behaviors of netizens, for example, by polarizing opinions, breeding hostility, or triggering mimicry and contagion of undesirable verbal interactions. Such cognitive or behavioral consequences of incivility are beyond the scope of this study. Therefore, our findings on the diffusion of profanity should *not* be overinterpreted as an ephemeral impact of incivility on the functioning of discursive participation.

Another limitation is that we operationalized profane speech somewhat narrowly by using a compiled list of swearing words and expressions. Although we manually verified that the swearing posts included in the diffusion analyses were representative of profanity, our narrow definition of profanity contrasted with the vague and manifold notion of vulgarity used by the Chinese authorities. Furthermore, even though we collected Weibo posts in real time filtered by specified keywords in order to capture profane posts immediately after they were published, there is a possibility that some posts had already been deleted before crawling and thus was not included in our dataset. Sina Weibo and other popular Chinese social networks are well known for exercising automated review and real-time manned monitoring of discussions and searches. As such, it is not verifiable as to how these censorship practices influenced this study's data collection and the subsequent analysis.

Also, this study only examined textual cues. As shown by the case of the grass-mud horse video, Chinese digital space includes a vast volume of multimedia content, the diffusion of which requires future research. Despite limitations, this study is one of the few that examined the diffusion process of profanity in a non-Western context. While we only focused on profane speech, future research may broaden the analysis by taking other types of uncivil behaviors online into consideration.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported in this article was supported by a General Research Fund grant (HKBU 12632816) of the Research Grant Council of Hong Kong, the Initiation Grant (Project No. RC-FNRA-IG-18-19-04) of Hong Kong Baptist University, and the Dean's Research Grant of Walter Cronkite School of Journalism and Mass Communication at Arizona State University.

ORCID iDs

Yunya Song https://orcid.org/0000-0001-5159-1689

K Hazel Kwon https://orcid.org/0000-0001-7414-6959

Note

1. A one-hot representation is a transformation of categorical features (in our example, words) into a format that can improve the predictive performance of machine learning algorithms, compared with using, for example, dummy variables. Each category is represented by a binary vector (a sequence of 0s and 1s), with a value of 1 for the target category and 0s for the other categories.

References

Anderson AA and Huntington HE (2017) Social media, science, and attack discourse: how Twitter discussions of climate change use sarcasm and incivility. *Science Communication* 39(5): 598–620.

Anderson AA, Yeo SK, Brossard D, et al. (2016) Toxic talk: how online incivility can undermine perceptions of media. *International Journal of Public Opinion Research* 30(1): 156–168.

Antoci A, Delfino A, Paglieri F, et al. (2016) Civility vs. incivility in online social interactions: an evolutionary approach. *PLoS ONE* 11(11): e0164286.

Arefi MN, Pandi R, Tschantz MC, et al. (2019) Assessing post deletion in Sina Weibo: multi-modal classification of hot topics. *arXiv preprint arXiv:1906.10861*. Available at: https://arxiv.org/abs/1906.10861

Boatright RG, Shaffer TJ, Sobieraj S, et al. (eds) (2019) *A Crisis of Civility? Political Discourse and its Discontents*. New York: Routledge.

Berger J and Milkman KL (2012) What makes online content viral? *Journal of Marketing Research* 49(11): 192–205.

Cavazza N and Guidetti M (2014) Swearing in political discourse: why vulgarity works. *Journal of Language and Social Psychology* 33(5): 537–547.

Cho D, Kim S and Acquisti A (2012, January) Empirical analysis of online anonymity and user behaviors: the impact of real name policy. In: *2012 45th Hawaii international conference on system sciences*, pp. 3041–3050. IEEE.

Cho D and Kwon KH (2015) The impacts of identity verification and disclosure of social cues on flaming in online user comments. *Computers in Human Behavior* 51: 363–372.

Coleman G (2014) *Hacker, Hoaxer, Whistleblower, Spy: The Many Faces of Anonymous*. New York: Verso books.

CNNIC (2019) The 44th China statistical report on Internet development. Available at: http://www.cnnic.cn/hlwfzyj/hlwxzbg/hlwtjbg/201908/P020190830356787490958.pdf

Diakopoulos N and Naaman M (2011) Towards quality discourse in online news comments. In: *Proceedings of the ACM 2011 conference on computer supported cooperative work*, Hangzhou, China, 19–23 March, pp. 133–142. New York: ACM.

de Seta G (2018) (Un) civil society in digital China| wenming bu wenming: the socialization of incivility in postdigital China. *International Journal of Communication* 12: 2010–2030.

Fägersten KB (2017) The role of swearing in creating an online persona: the case of YouTuber PewDiePie. *Discourse, Context & Media* 18: 1–10.

Ferrara E and Yang Z (2015) Measuring emotional contagion in social media. *PLoS ONE* 10(11): e0142390.

Ferree MM, Gamson WA, Gerhards J, et al. (2002) Four models of the public sphere in modern democracies. *Theory and Society* 31(3): 289–324.

Goel S, Anderson A, Hofman J, et al. (2016) The structural virality of online diffusion. *Management Science* 62(1): 180–196.

Gonzales AL, Hancock JT and Pennebaker JW (2010) Language style matching as a predictor of social dynamics in small groups. *Communication Research* 37(1): 3–19.

Hardaker C (2010) Trolling in asynchronous computer-mediated communication: from user discussions to academic definitions. *Journal of Politeness Research* 6: 215–242.

Hou R (2018) *The booming industry of Chinese state internet control*. *Hong Kong Free Press*, 30 November. Available at: https://www.hongkongfp.com/2018/11/30/booming-industry-chinese-state-internet-control/

Hwang H, Kim Y and Huh CU (2014) Seeing is believing: effects of uncivil online debate on political polarization and expectations of deliberation. *Journal of Broadcasting & Electronic Media* 58(4): 621–633.

Jay T (2000) *Why We Curse: A Neuro-Psycho-Social Theory of Speech*. Amsterdam: John Benjamins Publishing Company.

Jay T and Janschewitz K (2008) The pragmatics of swearing. *Journal of Politeness Research* 4: 267–288.

Jamieson KH, Volinsky A, Weitz I, et al. (2017) The political uses and abuses of civility and incivility. *The Oxford Handbook of Political Communication:* 205–218.

Jefferis VE, van Baaren R and Chartrand TL (2003) *The Functional Purpose of Mimicry for Creating Interpersonal Closeness* (*Manuscript)*. Columbus, OH: The Ohio State University.

Jiang M and Esarey A (2018) Uncivil society in digital China: incivility, fragmentation, and political stability. *International Journal of Communication* 12: 1928–1944.

Katz E (1957) The two-step flow of communication: an up-to-date report on an hypothesis. *Public Opinion Quarterly* 21(1): 61–78.

Kasson JF (1990) *Rudeness & Civility: Manners in Nineteenth-Century Urban America*, vol. 2. New York: Macmillan.

Kennedy R (1998) Civility in the house of representatives. *American Prospect* 41: 84–91.

Kenski K, Coe K and Rains SA (2017) Perceptions of uncivil discourse online: an examination of types and predictors. *Communication Research*. Epub ahead of print 3 April. DOI: 10.1177/0093650217699933.

Kim Y and Herring SC (2018) Is politeness catalytic and contagious? Effects on participation in online news discussions. In: Proceedings of the 51st Hawaii international conference on system sciences, 3–6 January. Available at: https://core.ac.uk/download/pdf/143481085.pdf

King G, Pan J and Roberts ME (2013) How censorship in China allows government criticism but silences collective expression. *American Political Science Review* 107(2): 326–343.

Kwon KH and Cho D (2017) Swearing effects on citizen-to-citizen commenting online: a large-scale exploration of political versus nonpolitical online news sites. *Social Science Computer Review* 35(1): 84–102.

Kwon KH and Gruzd A (2017) Is offensive commenting contagious online? Examining public vs interpersonal swearing in response to Donald Trump's YouTube campaign videos. *Internet Research* 27(4): 991–1010.

Li H (2011) Parody and resistance on the Chinese Internet. In: Herold D and Marolt PW (eds) *Online Society in China*. London: Routledge, pp. 71–88.

Liang B and Lu H (2012) Fighting the obscene, pornographic, and unhealthy—an analysis of the nature, extent, and regulation of China's online pornography within a global context. *Crime, Law and Social Change* 58(2): 111–130.

Liu-Thompkins Y (2012) Seeding viral content: the role of message and network factors. *Journal of Advertising Research* 52(4): 465–478.

Liu Y, Jin X, Shen H, et al. (2017) Do rumors diffuse differently from non-rumors? A systematically empirical analysis in Sina Weibo for rumor identification. In: *Pacific-Asia conference on knowledge discovery and data mining*. Cham: Springer, pp. 407–420.

Mak BCN and Lee C (2015) Swearing is e-business: expletives in instant messaging in Hong Kong workplaces. In: Darics E *(ed.) Digital Business Discourse*. London: Palgrave Macmillan, pp. 122–141.

Muddiman A (2017) Personal and public levels of political incivility. *International Journal of Communication* 11: 3182–3202.

Muddiman A and Stroud NJ (2017) News values, cognitive biases, and partisan incivility in comment sections. *Journal of Communication* 67(4): 586–609.

Nahon K and Hemsley J (2013) *Going Viral*. Cambridge: Polity Press.

O'Sullivan PB and Flanagin AJ (2003) Reconceptualizing "flaming" and other problematic messages. *New Media & Society* 5(1): 69–94.

Song Y and Wu Y (2018) Tracking the viral spread of incivility on social networking sites: the case of cursing in online discussions of Hong Kong–Mainland China conflict. *Communication and the Public* 3(1): 46–61.

Song Y and Xu R (2019) Affective ties that bind: investigating the affordances of social networking sites for commemoration of traumatic events. *Social Science Computer Review* 37(3): 333–354.

Song Y, Dai XY and Wang J (2016) Not all emotions are created equal: expressive behavior of the networked public on China's social media site. *Computers in Human Behavior* 60: 525–533.

Stieglitz S and Dang-Xuan L (2013) Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. *Journal of Management Information Systems* 29(4): 217–248.

Strachan J and Wolf M (2012) Calls for civility: an invitation to deliberate or a means of political control? In: Shea D and Fiorina M (eds) *Can We Talk? The Rise of Rude, Nasty, Stubborn Politics*. New York: Pearson, pp. 41–52.

Stryker R, Conway BA and Danielson JT (2016) What is political incivility? *Communication Monographs* 83(4): 535–556.

Suler J (2004) The online disinhibition effect. *Cyberpsychology & Behavior* 7(3): 321–326.

Tang L and Yang P (2011) Symbolic power and the Internet: the power of a "horse." *Media, Culture & Society* 33(5): 675–691.

Taylor PJ and Thomas S (2008) Linguistic style matching and negotiation outcome. *Negotiation and Conflict Management Research* 1(3): 263–281.

Vandebosch H and Van Cleemput K (2009) Cyberbullying among youngsters: profiles of bullies and victims. *New Media & Society* 11(8): 1349–1371.

Vosoughi S, Roy D and Aral S (2018) The spread of true and false news online. *Science* 359(6380): 1146–1151.

Wang SS (2012) China's Internet lexicon: symbolic meaning and commoditization of Grass Mud Horse in the harmonious society. *First Monday* 17(1). DOI: 10.5210/fm.v17i1.3758.

Wu G (2009) In the name of good governance: E-government, Internet pornography and political censorship in China. In: Zhang X and Zheng Y (eds) *China's Information and Communications Technology Revolution*. London: Routledge, pp. 80–97.

Yang G (2018) (Un)civil society in digital China| demobilizing the emotions of online activism in china: a civilizing process. *International Journal of Communication* 12: 21.

Yoganarasimhan H (2011) Impact of social network structure on content propagation: a study using YouTube data. *Quantitative Marketing and Economics* 10(1): 111–150.

Zhang Y and Wallace B (2015) A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820.

Author biographies

Yunya Song is an associate professor in the School of Communicationsm and Director of AI and Media Research Lab, Hong Kong Baptist University. Her research on new media and political communication has appeared in, among other journals, *International Journal of Press/Politics, Computers in Human Behavior*, and *Social Science Computer Review*. She is the Co-Editor of *The Evolving Landscape of Media and Communication in Hong Kong* (City University of Hong Kong Press, 2018).

K Hazel Kwon is an associate professor in the Walter Cronkite School of Journalism and Mass Communication at Arizona State University. Her publications have appeared in multiple journals, including the *Journal of Computer-Mediated Communication, CyberPsychology, Behavior & Social Networking, Computers in Human Behaviors, American Behavioral Scientists*, and *Social Science Computer Review*.

Jianliang Xu is a professor in the Department of Computer Science at Hong Kong Baptist University, where he leads the Database Research Group. His current research interests include big data management, data security and privacy, and blockchain technology. With an h-index of 45, he has published more than 200 technical papers in these areas, most of which appeared in leading journals and conferences including SIGMOD, PVLDB, ICDE, TODS, TKDE, and VLDBJ.

Xin Huang is an assistant professor in the Department of Computer Science at Hong Kong Baptist University. He received the PhD degree from the Chinese University of Hong Kong in 2014. His research interests mainly focus on graph data management and mining.

Shiying Li studies in the Department of Computer Science at Brown University.