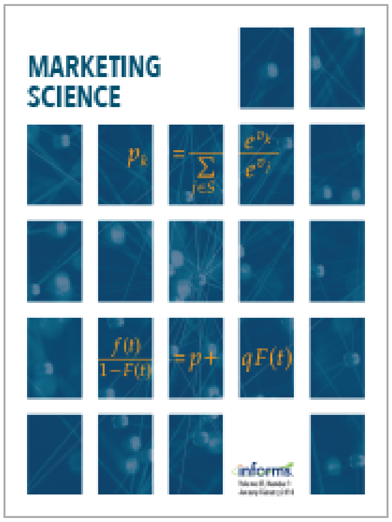


Marketing Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Does Online Word of Mouth Increase Demand? (And How?) Evidence from a Natural Experiment

Stephan Seiler, Song Yao, Wenbo Wang

informs
http://pubsonline.informs.org/journal/mksc/

MARKETING SCIENCE
Vol. 36, No. 6, November–December 2017, pp. 838–861
ISSN 0732-2399 (print), ISSN 1526-548X (online)

Does Online Word of Mouth Increase Demand? (And How?) Evidence from a Natural Experiment

Stephan Seiler,[a] Song Yao,[b] Wenbo Wang[c]

[a] Stanford University, Stanford, California 94305; [b] Carlson School of Management, University of Minnesota, Minneapolis, Minnesota 55455; [c] Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong

Contact: sseiler@stanford.edu (SS); syao@umn.edu, http://orcid.org/0000-0002-0339-9649 (SY); wenbowang@ust.hk (WW)

Received: November 20, 2015
Revised: July 28, 2016; February 21, 2017
Accepted: February 28, 2017
Published Online in Articles in Advance: September 28, 2017

https://doi.org/10.1287/mksc.2017.1045

Abstract. We leverage a temporary block of the Chinese microblogging platform Sina Weibo due to political events to estimate the causal effect of online word-of-mouth content on product demand in the context of TV show viewership. Based on this source of exogenous variation, we estimate an elasticity of TV show ratings (market share in terms of viewership) with respect to the number of relevant comments (comments were disabled during the block) of 0.016. We find that more postshow microblogging activity increases demand, whereas comments posted prior to the show airing do not affect viewership. These patterns are inconsistent with informative or persuasive effects and suggest complementarity between TV consumption and anticipated postshow microblogging activity.

History: Avi Goldfarb served as the senior editor and Vishal Singh served as associate editor for this article.

Funding: This research was partially funded by the Hong Kong Research Grants Council Early Career Scheme [Grant 699213], Hong Kong University of Science and Technology Institute for Emerging Market Studies [Research Grant 16BM07], and the McManus Research Grant from the Kellogg School of Management, Northwestern University.

Supplemental Material: Data and the online appendix are available at https://doi.org/10.1287/mksc.2017.1045.

Keywords: microblogging • advertising • social media • word of mouth • natural experiment

1. Introduction

With the advent of web 2.0, the exchange of information between consumers has become increasingly public, and word of mouth (WOM), which used to be confined to a small circle of contacts in the offline world, has tremendously increased in visibility in the online context. The proliferation of WOM via social media platforms such as Facebook and Twitter raises the question of how strongly brand-related online WOM can affect product sales. One obstacle to answering this question is that causal inference is particularly difficult in the realm of online WOM because firms are not directly in control of the amount of WOM. Therefore, randomized field experiments, which have increased in prominence in studies of advertising effectiveness (Lewis and Reiley 2014, Blake et al. 2015, Sahni 2015, Gordon et al. 2016), are difficult to implement in the case of WOM.

In this paper, we leverage a natural experiment, the temporary shutdown of the most popular Chinese microblogging outlet Sina Weibo because of political events, to estimate the causal effect of online WOM on product demand in the context of TV show viewership. Based on this source of exogenous variation, this paper makes the following contributions. First, we quantify the *causal* impact of WOM on product sales. Second, we analyze and identify the behavioral mechanism underlying the effect with a particular focus on complementarities between microblogging and TV consumption, a mechanism that has received little attention in previous work.[1]

Two key findings emerge from our analysis. First, we find an elasticity of TV viewership with respect to microblogging activity of 0.016, which is substantially lower than most estimates in the previous literature. For instance, Sonnier et al. (2011), Dhar and Chang (2015), and Liu (2015) estimate WOM elasticities between 0.59 and 1.04 across a variety of markets.[2] We suspect that part of this discrepancy can be attributed to the fact that previous papers were limited in their ability to deal with endogeneity issues, because of the absence of (quasi-)experimental variation. Our estimated elasticity is slightly smaller than elasticities typically estimated for TV advertising of around 0.03 (Gordon and Hartmann 2013, Shapiro 2016, Tuchman 2016), rather than substantially larger, as other estimates would suggest. This finding is timely because of the recent increase in marketing spending on social media and a general belief among marketing practitioners that WOM can be a highly effective way to reach customers. For example, according to a study by the American Marketing Association, 64% of marketing executives believe WOM is the most effective form of marketing (PR Newswire 2013). Our findings caution against overestimating the effectiveness of WOM relative to other marketing channels.[3]

The second key contribution of this paper is to identify the behavioral mechanism by which microblogging affects consumers' TV viewing decisions. We do so by exploiting detailed information on the timing of microblogging activity as well as an extensive content analysis. In a first step, we distinguish two possible behavioral channels that could drive the effect of WOM on product demand. WOM could serve as a complementary activity to product consumption, and thus its presence increases demand because it enhances the utility derived from watching the show. Alternatively, WOM could affect demand in a similar fashion as traditional advertising by informing or persuading people to watch a specific show. Exploiting differences in the timing of microblogging activity, we find that microblogging activity *after a show* has aired is the primary driver of viewership, whereas the amount of activity prior to the show does not have an impact on viewership. This pattern suggests the complementarity between TV viewing and the consumption of (postshow) microblogging content increases TV ratings (market share in terms of viewership). Instead, if informative or persuasive effects were important, we would expect preshow microblogging activity to affect ratings. Next, we investigate which type of postshow activity most strongly affects ratings and find that microblogging activity expressing sentiment has the strongest effect. Interestingly, both positive and negative sentiment affect ratings positively, suggesting that an engaging (and potentially controversial) postshow debate is the key driver of TV show ratings.

Our analysis proceeds in the following steps. First, we provide details on the reasons behind the shutdown of the microblogging platform Sina Weibo and show the political events that triggered the shutdown are unlikely to have had any direct effect on TV show viewership. The reason behind the temporary block was the defection of a prominent government official in early 2012, which led to a series of related events over a period of three months. Roughly midway through this time period, the Chinese government limited the functionality of Sina Weibo for three days. The censorship did not block Sina Weibo entirely, but only disabled the commenting function. In other words, users were still able to post tweets but were unable to comment on those tweets. Using Google trends and Baidu search-volume data as well as counts of news stories pertaining to the scandal, we show people's interest in the scandal increased at three distinct points in time that were associated with major events of the scandal. None of these events, however, coincided with the Weibo censorship, and the spikes in interest did not lead to a change in TV viewership. These patterns provide our first piece of evidence that a direct effect of the block is unlikely.

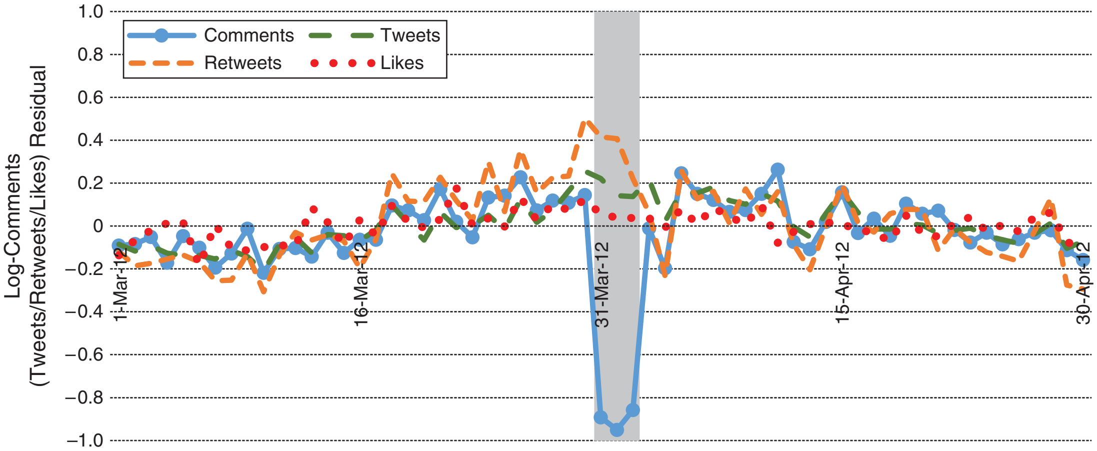
Second, we use a series of difference-in-differences (DiD) regressions to further rule out a direct effect of the block or other contemporaneous events on TV ratings. We show TV show ratings dropped in mainland China, but not in Hong Kong, where Twitter rather than Sina Weibo is used. Furthermore, within mainland China, ratings decreased more in cities with higher Weibo usage. Finally, we find the block only affected shows with a high level of microblogging activity, whereas shows with little to no activity did not experience any change in ratings during the censorship. These differential effects across geographies and shows provide a second piece of evidence that the block affected TV ratings via the channel of reduced Weibo comments.

Having established that the block affected ratings only via its impact on online WOM, we proceed in a third step to estimate the elasticity of ratings with respect to online WOM. To obtain this elasticity estimate, we run an instrumental variable (IV) regression in which we regress episode-level ratings on the number of comments posted on the day the episode aired as well as show and weekday fixed effects. We use a dummy for the time period of the block as an instrument for the number of comments to isolate the part of the variation driven by the natural experiment.

We note that the block affects microblogging activity in two ways that are relevant to viewing decisions. It reduces both preshow comments and anticipated postshow comments. Our IV regression does not distinguish between these channels. For simplicity, we use the statement "comments cause ratings" to refer to ratings being causally affected by either preshow comments or anticipated postshow comments. In Section 6 we disentangle these two mechanisms.

The key variation the IV regression exploits is illustrated by the graphs provided in Figures 1 and 2, which display the time series of the movement in Weibo comments[4] and show ratings, respectively. Figure 1 plots the average daily residuals from a regression at the episode level of (log) comments onto show (and weekday) fixed effects, thus isolating the time series variation in comments. The graph highlights the decrease in comments during the block, which occurred March 31 to April 2 and is indicated with a gray band. Figure 2 presents a similar graph for TV show ratings, which decreased during the time period of the Sina Weibo block relative to the typical rating variation over time. Our IV regression relates the drop in comments to the drop in ratings during the block. As a further piece of evidence for the ratings decrease during the block, we also compute the change in ratings for the top 10

Figure 1. (Color online) The Evolution of Sina Weibo Activity Over Time



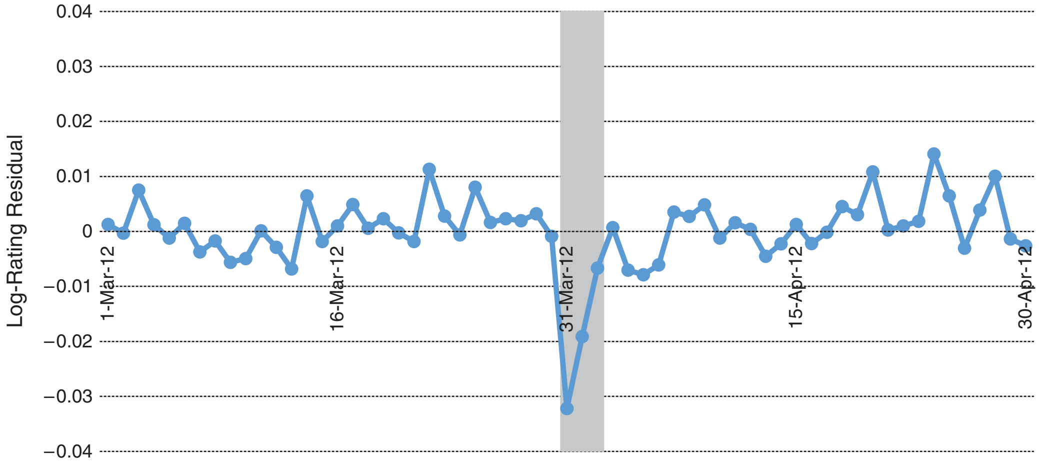
Notes. The graph plots the average daily values of the residuals from a set of regressions of log(comments), log(tweets), log(retweets) and log(likes) on show and weekday fixed effects. The gray band indicates the three days of the block.

shows (based on their average ratings) at the individual show level. We find all 10 shows experienced lower ratings during the censorship, and the decrease is statistically significant for most shows. Detailed results are reported in Table A1 in the online appendix.

We note the decrease in ratings during the block is modest in magnitude. In Figure 2, we therefore control more rigorously for factors that lead to volatility in ratings over time than in most of our regressions. These controls allow us to isolate the ratings reduction during the censorship more clearly.[5] As a comparison, we present a similar graph that controls only for show and weekday fixed effects (as we do in all of our regressions) in Figure A1 in the online appendix. Volatility of ratings outside of the block is higher, and hence the decrease during the block is less visually salient (but still statistically significant).[6] This lack of salience might not be surprising because, similar to the impact of advertising (see Lewis and Rao 2015), we would not expect WOM variation to have dramatic effects on ratings. Although none of our regressions suffer from a lack of sufficient statistical power, we nevertheless emphasize that in a setting with a small effect magnitude like ours, it is reassuring that we are able to consistently detect the effect across a series of DiD specifications. We also probe the robustness of our results to different ways of constructing standard errors and find the statistical power of our results to be unaffected.

In a fourth and final step, we explore the mechanism underlying the impact of WOM on TV show ratings. As outlined above, we investigate heterogeneity in the effect of comments on ratings as a function of their

Figure 2. (Color online) The Evolution of Ratings Over Time



Notes. The graph plots the average daily values of the residuals from a regression of log(rating) on show/weekday pair fixed effects and time-varying controls. The gray band indicates the three days of the block. Time-varying controls include holidays and a control for the duration of the world team table tennis championship.

timing (before or after the show) as well as their content, and find postshow comments expressing (positive or negative) sentiment are the primary driver of viewership. We interpret these data patterns as evidence that the complementarity of TV and anticipated postshow microblogging consumption is the main channel through which WOM affects demand in our setting.

Our paper contributes to various streams of literature. First, in terms of the substantive questions addressed, our paper relates to the growing literature on social media as a marketing channel. This general area can be further divided into studies of firms' own advertising as well as user-generated content pertaining to specific brands or products.[7] This distinction is sometimes referred to as paid/owned versus earned media. Within the former category, Tucker (2014) shows advertising based on personal information can increase clicks, Aral and Walker (2014) analyze how firms can generate viral marketing strategies on social networks, and Lambrecht et al. (2017) study whether Twitter promotions are more effective on early trend adopters. Petrova et al. (2016) study the impact of Twitter activity on donations to political campaigns. Gong et al. (2017) conduct a field experiment to measure microblogging's effect on demand, by varying the level of tweets generated by the firm.

Within the realm of social media marketing that is not under the firm's control, one can further distinguish studies of customer reviews as well as the analysis of online conversations about brands, which our setting falls into. This distinction is helpful because the mechanisms underlying both types of WOM are distinct. Reviews are typically written after consumption and do not involve a conversation between users. Therefore, the mechanism pertaining to consumption complementarity that we investigate is specific to online conversations and less likely to apply to reviews. A series of papers uses quasi-experimental methods to understand the impact of the valence and volume of reviews on sales. Chevalier and Mayzlin (2006) and Zhu and Zhang (2010) use DiD approaches across platforms, and Luca (2016) and Anderson and Magruder (2012) employ regression discontinuity designs. Chintagunta et al. (2010) base their analysis on the sequential roll-out of movies across different markets.

With regard to papers analyzing the impact of online conversations, we are not aware of any that use experimental or quasi-experimental methods. To our knowledge, our paper is the first to leverage a natural experiment to estimate the effect of WOM in this context. The bulk of the literature on (nonreview) WOM uses some form of dynamic panel model to relate sales and WOM in different time periods to each other (Godes and Mayzlin 2004, Trusov et al. 2009, Villanueva et al. 2008). Although these approaches allow us to better understand the temporal correlation patterns of WOM and sales, the nature of the variation is unlikely to recover a causal effect. These methodological differences manifest themselves in the estimated elasticity, which we find to be significantly lower than estimates found in the prior literature.

In terms of the identification strategy, our approach is related to a small set of papers that measure advertising effectiveness by exploiting quasi-experimental variation. Hartmann and Klapper (2017) exploit across-market variation in TV advertising exposure during the Super Bowl. Shapiro (2016) and Tuchman (2016) use ad-exposure variation in the vicinity of borders between neighboring TV markets to study antidepressant and e-cigarette advertising. Sinkinson and Starc (2015) use the crowding out of regular ads by political campaign ads as a source of exogenous variation. Our study similarly exploits quasi-experimental variation, in our case provided by the censorship event on Sina Weibo.

In terms of the behavioral mechanism underlying our findings, our paper relates to the work of Becker and Murphy (1993), who propose a model of complementarity between product consumption and advertising, and Tuchman et al. (2017), who empirically test this model. In our setting, consumption complementarity, arising from consumers' social interactions rather than advertising, plays a key role. In a related paper, Gilchrist and Sands (2016) show that social interactions increase movie viewership because of consumers wanting to enjoy a "shared experience" with others.

Finally, one paper that shares several features with ours is Lovett and Staelin (2016) who also investigate the impact of WOM on TV viewership using a structural model and detailed data from one TV series. Their paper estimates a WOM elasticity that is similar in magnitude to ours, even though it is not based on a specific source of exogenous variation. They also find that WOM is less effective than advertising. In terms of mechanism, Lovett and Staelin (2016) highlight the importance of "enhanced enjoyment" of live TV consumption that arises from viewers interacting after the show.

The remainder of this paper is structured as follows. Sections 2 and 3 outline the data and provide details on the nature of the Sina Weibo block. In Section 4, we present results from a set of DiD regressions. In Section 5, we implement an IV regression to measure the elasticity of viewership with respect to the number of comments on Sina Weibo tweets. In Sections 6 and 7, we investigate the behavioral mechanism driving our results and dynamic effects of microblogging on demand. Finally, we provide some concluding remarks.

2. Data and Descriptive Statistics

We rely on two separate sources of data in this paper. First, we use detailed episode-level data on TV viewership (i.e., ratings) across a large set of shows, as well as geographic locations in China. Second, we assemble a unique data set of microblogging activity by scraping Sina Weibo content for the set of shows that appear in the TV data. We outline in detail the two data sets and provide descriptive statistics.

2.1. TV Ratings Data

We obtain data on viewership for a large set of TV shows in mainland China and Hong Kong from CSM Media Research, the leading TV-rating data provider in China. The data are reported at the episode-city level. Each episode belongs to a specific show, and each show may comprise multiple episodes if the show is a series. Because of constraints imposed by the data provider, we do not observe the full universe of shows. Instead, we work with data from an extensive subset of shows and cities. Specifically, we select a set of 24 major cities in mainland China and Hong Kong (see Figure A2 in the online appendix for the cities' locations). For each city (except Hong Kong), the set of shows is identical. Next, we select the top 20 national channels based on their market share in mainland China, which covers roughly 80% of the mainland market.[8] In Hong Kong, we obtain data for all six local channels. For each city and channel combination, we collect ratings for all TV shows that ran between 6 P.M. and 12 A.M. every day from March 1 to April 30, 2012. We choose these two months based on the timing of the Sina Weibo block, which took place from March 31 to April 2, 2012.[9] Our time window therefore covers roughly a month before and a month after the block.

In a final step, we further narrow the set of shows down to the ones that provide relevant variation for our analysis. Specifically, all of our later regressions will analyze the change in viewership for a given show during the block relative to episodes of the same show that aired before or after the block. We therefore focus on shows for which we observe multiple episodes and which aired at least one episode during the block and one episode before or after the block. This selection leaves us with a total of 166 shows in mainland China. Because some shows were aired (usually on different dates and times) on multiple channels, we have 193 show/channel pairs. For Hong Kong, we have 112 shows and 132 show/channel pairs. In all of our analysis, a show/channel combination constitutes the cross-sectional unit; going forward, we simply refer to such a combination as a "show."

Because the set of mainland shows is identical across cities, we base most of our analysis on the average market share of each episode across all cities in mainland China. When computing the average market share, we weigh the observations from the different cities by their population size. Table 1 presents descriptive statistics for the aggregated data on the 193 mainland shows for which we observe a total of 7,899 individual episodes. Our main outcome measure for each episode is the episode's rating, that is, its market share in terms of viewership (measured from 0 to 100 rating points). In the first two rows of Table 1, we report ratings as well as log-ratings across all 7,899 episodes. The average episode has a rating of 0.434, but the distribution is relatively skewed with a large standard deviation and a maximum rating of 4.166 in our sample.[10] The shows in our sample come from a variety of genres ranging from TV serials and reality shows to reporting of current events and shows geared toward children. In Table A2 in the online appendix, we report the rating distribution across shows for each genre separately.

We also decompose the variation in ratings into the across-show variation as well as the time-series variation within shows by computing the residuals from a regression of ratings onto show fixed effects. The standard deviation of the residuals is reported in the second to last column of Table 1. We provide the same descriptive statistics for shows in Hong Kong in Table A3 in the online appendix. Finally, we note that many shows in China are broadcast at a high frequency, and the average interval between two consecutive episodes is 1.48 days. We therefore have a substantial panel dimension in our data. For the two months of data in our sample, we observe 7,899 episodes for 193 shows and hence have about 40 observations per show.

2.2. Microblogging Data

Our second data set measures the amount of activity related to each TV show on Sina Weibo, which is the primary microblogging website in China, with 61 million daily active users and around 100 million daily tweets.[11] On Sina Weibo, users can engage in four different types of activities: tweeting, retweeting, liking, and commenting. Commenting is the type of activity primarily affected by the block. It allows a user to respond to an existing tweet with some content of her own (a "comment"). All comments on a given tweet are listed below that tweet, forming a thread of interactive discussion among users who are interested in the subject of that tweet. These users can read and reply to each other's content in that thread, even though they do not necessarily follow each other. The conversational nature of comments distinguishes it from other types of activity (e.g., tweets, retweets, and likes) and is likely to play a role in the behavioral mechanism we identify later. Figure 3 displays an example of a tweet together with a series of comments pertaining to the tweet.

We obtained the microblogging data by scraping the Sina Weibo website. Specifically, for all shows contained in our data, we scraped every tweet mentioning the show during March and April of 2012. We further collected the number of comments on each tweet,[12] as

Table 1. Descriptive Statistics: TV Ratings and Sina Weibo Activity

	Mean	S.D.	90th perc.	95th perc.	Max.	S.D. (time series only)	Obs.
Ratings							
Rating	0.434	0.483	0.941	1.225	4.166	0.158	7,899
Log rating	0.322	0.260	0.663	0.800	1.642	0.090	7,899
Microblogging							
Comments	11,311	76,757	345	6,670	2,506,875	48,962	7,899
Log comments	2.18	3.03	5.84	8.81	14.73	1.25	7,899
Tweets	2,294	11,047	125	4,083	109,923	5,659	7,899
Retweets	68,192	960,259	622	21,813	30,142,880	921,605	7,899
Likes	386	9,640	1	27	487,849	9,290	7,899
	Mean	S.D.	Min	10th perc.	90th perc.	Max.	Obs.
Show-level comments							
All shows	14,406	83,159	0	0	416	786,398	193
TV series	24,505	73,876	1	7	37,534	257,320	23
Reality shows	48,632	169,142	0	0	111,139	786,398	40
Children's shows	26	69	0	0	45	279	16
News	5,359	21,261	0	0	260	88,665	50
Other shows	49	152	0	0	110	1,052	64
Established shows	16,158	93,862	0	0	281	786,399	137
New shows	10,120	48,269	0	0	1,333	257,320	56
Rerun	59	107	0	0	176	416	22
Current show	16,252	88,206	0	0	571	786,399	171
Daily frequency	7,060	35,994	0	0	279	257,320	118
Less than daily frequency	25,965	125,192	0	0	1,302	786,399	75

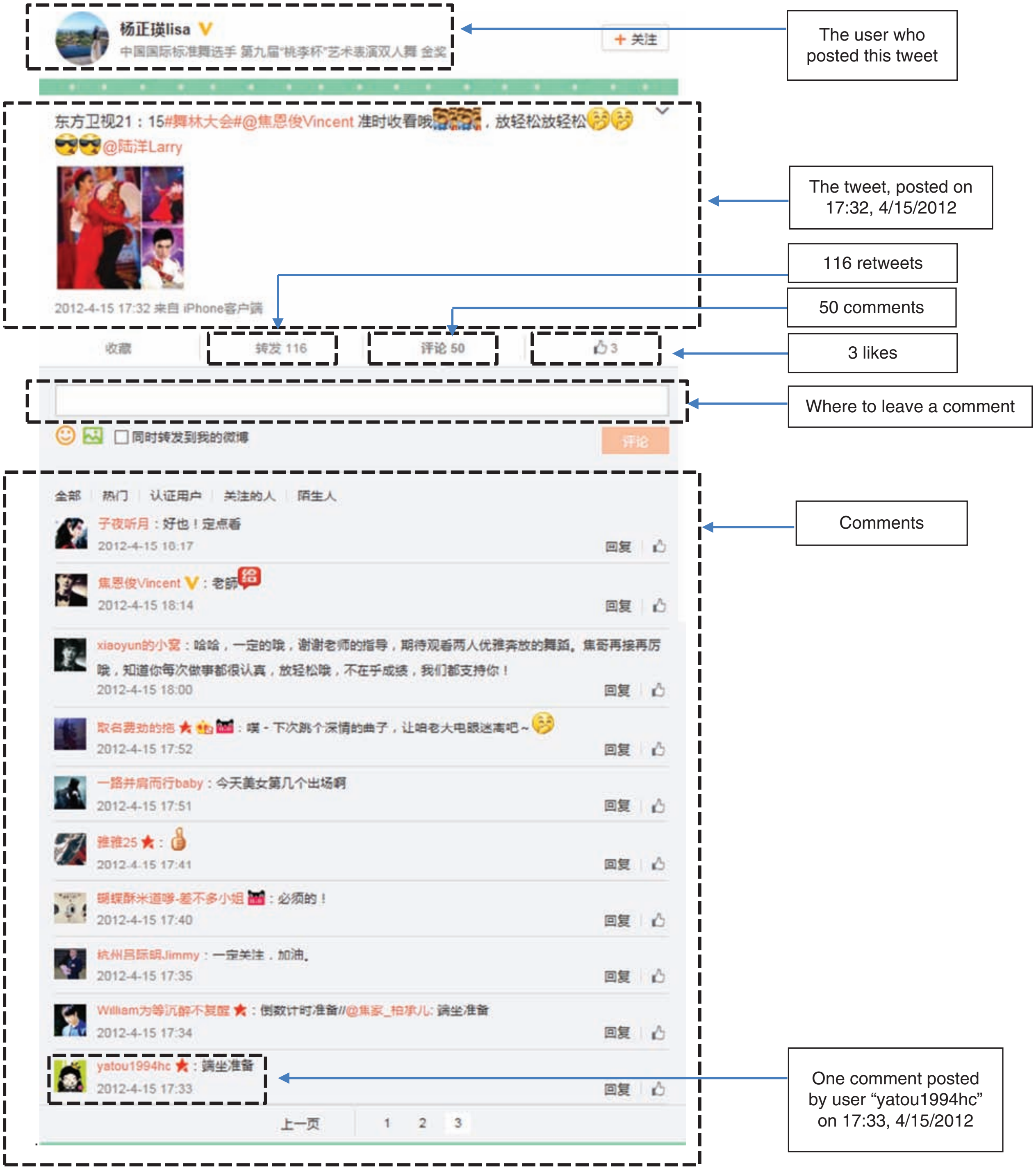
Note. The unit of observation is an episode in the top panel and a show in the bottom panel.

well as the number of retweets and likes.[13] To relate the amount of Sina Weibo activity to a particular episode, we calculate the number of relevant tweets, retweets, comments, and likes that users posted on the day that a particular episode aired.[14] We note this definition includes microblogging activity both before and after the show aired. Based on a cursory check of the content of tweets and comments, we found a calendar day to be a good approximation for the time window that delineates content pertaining to a specific episode. Preshow tweets typically contain content regarding consumers' anticipation of the upcoming episode, whereas postshow tweets contain discussion of the episode that aired earlier in the day. We return in more detail to the timing of activity when investigating the mechanism by which microblogging affects demand in Section 6.

We report descriptive statistics for the different types of microblogging activity in the lower portion of the first panel in Table 1. First, we note that comments, which are the type of activity primarily affected by the block, are frequently used, and for the average episode, the number of comments is about five times larger than the number of tweets. In terms of other types of user activity, we observe a large number of retweets related to the shows in our sample. The "like" feature is used infrequently. These four types of user activities represent the exhaustive set of options for participating on Sina Weibo as a user. Although our primary focus will be on comments and their decrease during the block, we also track other types of activity and later assess whether the block indirectly affected them. In terms of the distribution of comments across episodes, we find a highly skewed distribution. The average number of comments per episode in our data is equal to 11,311, but some episodes in the right tail of the distribution are mentioned in a substantially larger number of comments. Similar patterns also hold for tweets, retweets, and likes.

To explore which show characteristics correlate with the amount of activity on Sina Weibo, we report the across-show distribution of comments for specific subsets of shows in the bottom panel of Table 1. The unit of observation is a show (rather than an episode), and for each show, we compute the average number of comments per episode, excluding any episodes that aired during the block. We first decompose the comment distribution by show genre. In our sample, we identify four major categories of TV shows, namely, TV series, reality shows, news/reporting on current events, and children's shows. The remaining shows belong to a variety of genres such as weather reports, finance-related shows, historical documentaries, and so on, and we lump them together in a residual category. TV series and reality shows garner substantially more comments per episode than children's shows and shows in the "other" category. News reporting also receives relatively high activity, largely because consumers debate

Figure 3. (Color online) Example of a Tweet with Comments



杨正瑛lisa V
中国国际标准舞选手 第九届"桃李杯"艺术表演双人舞 金奖
+ 关注

东方卫视21：15#舞林大会#@焦恩俊Vincent 准时收看哦，放轻松放轻松 @陆洋Larry

2012-4-15 17:32 来自 iPhone客户端

收藏 | 转发 116 | 评论 50 | 3

□ 同时转发到我的微博 评论

全部 | 热门 | 认证用户 | 关注的人 | 陌生人

子夜听月：好也！定点看
2012-4-15 18:17 回复

焦恩俊Vincent V：老師
2012-4-15 18:14 回复

xiaoyun的小窝：哈哈，一定的哦，谢谢老师的指导，期待观看两人优雅奔放的舞蹈。焦哥再接再厉哦，知道你每次做事都很认真，放轻松哦，不在乎成绩，我们都支持你！
2012-4-15 18:00 回复

取名费劲的拖 ★ ：噗 - 下次跳个深情的曲子，让咱老大电眼迷离吧 ~
2012-4-15 17:52 回复

一路并肩而行baby：今天美女第几个出场啊
2012-4-15 17:51 回复

雅雅25 ★ ：
2012-4-15 17:41 回复

蝴蝶酥米道哆-差不多小姐：必须的！
2012-4-15 17:40 回复

杭州吕际明Jimmy：一定关注，加油。
2012-4-15 17:35 回复

William为等沉醉不复醒 ★：倒数计时准备//@焦家_柏承儿: 端坐准备
2012-4-15 17:34 回复

yatou1994hc ★：端坐准备
2012-4-15 17:33 回复

上一页 1 2 3

The user who posted this tweet

The tweet, posted on 17:32, 4/15/2012

116 retweets

50 comments

3 likes

Where to leave a comment

Comments

One comment posted by user "yatou1994hc" on 17:33, 4/15/2012

Note. The displayed tweet is about a popular Chinese dance competition show, similar to the U.S. show *Dancing with the Stars*.

the news coverage and related political events on Sina Weibo.

We provide a similar decomposition for old versus new shows, where the latter are defined as shows that started airing during our sample period, as well as whether a show is a rerun or airing for the first time. Finally, we also analyze differences in activity across shows that air daily versus at a lower frequency. We find significant heterogeneity exists in activity within each genre, and the subset of shows in the new/old, rerun/first time, and daily/less frequent categories. Therefore, while observable show characteristics do predict the extent of microblogging activity, there are strong show-specific factors that drive the amount of Sina Weibo activity.

3. The Block

In this section, we provide details on the censorship of Sina Weibo and the events leading up to it. Other than providing background information on the source of exogenous variation we exploit, the time line of the scandal also helps us assess whether the block and the political events driving it might have affected TV viewership through any channel other than the reduction in microblogging activity. The timing of the censorship event thus provides a first piece of evidence in support of our exclusion restriction: The block affected TV ratings only via its impact on microblogging.

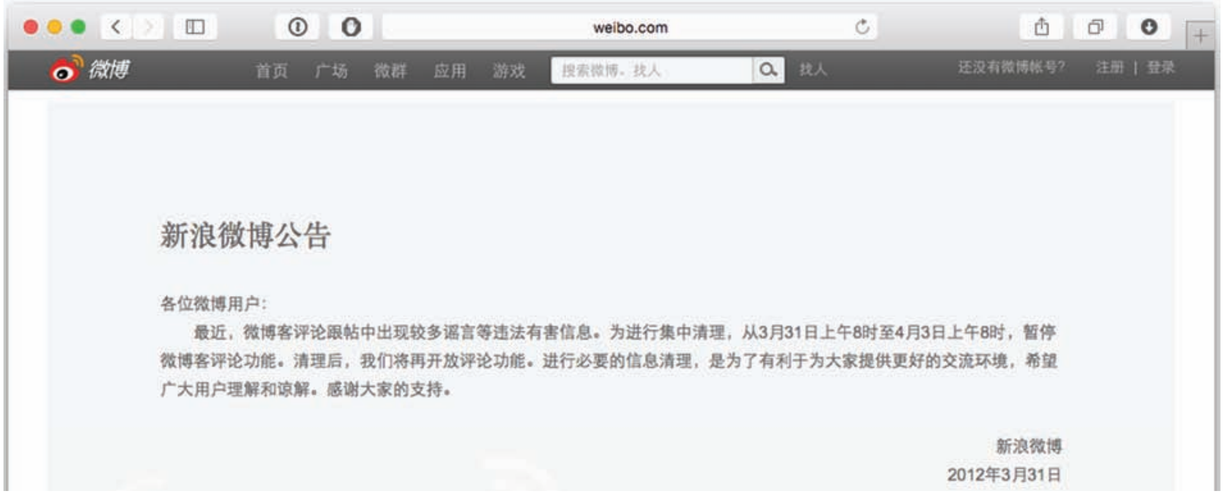
In February 2012, a political scandal erupted in China after Wang Lijun, a top government official in the city of Chongqing, defected to the U.S. Consulate. His superior, Bo Xilai, was later removed from his post and arrested. By March 2012, many rumors related to the political scandal appeared on the Internet, especially on social media websites. To remove such rumors, Sina Weibo announced early on March 31, 2012, that from 8 A.M. that day until 8 A.M. April 3, the microblogging platform would be partially blocked.[15] Specifically, the commenting function was disabled during that time period.[16] Figure 4 displays the announcement of the block that appeared on the Sina Weibo homepage and was visible to anybody using the platform. The statement reads in English as follows:

> To all Weibo users:
>
> Recently, there have been many harmful and illegal rumors that appeared in the comment section of Weibo. To effectively remove these rumors, from March 31 8 A.M. to April 3 8 A.M., we will suspend the comment function on Weibo. After removing the rumors, we will reopen the comment function. Such an action is necessary. It is for the purpose of creating a better environment for users' communications. We ask for your understanding and forgiveness. Thank you for your support.
>
> Sina Weibo
> March 31, 2012

Because the origins of the block were political events, the events that triggered the block might have had a direct effect on TV viewership. For instance, one could imagine consumers paid more attention to the political events during those three days and thus were less likely to watch TV. Such behavior would lead to a correlation between the time period of the block and TV viewership that is unrelated to the reduction in the number of comments. Two pieces of evidence, however, speak against the presence of such a direct effect. First, the political scandal unfolded over the course of about three months and involved several major events, none of which coincided with the block. Second, we find that during those major events, TV viewership remained unchanged, while it dropped during the Sina Weibo block. Therefore, interest in the political events does not seem to have affected viewership.

To establish the first point, we obtain data from the Chinese search engine Baidu regarding the number of search queries for the names of the two Chinese officials involved in the scandal: Wang Lijun and Bo Xilai. In Figure 5, we plot the time series of the daily number of searches for the two names from January to June 2012, where the three days of the Sina Weibo block are indicated in gray. The graphs display the time window of the entire series of events from early February to mid-April and feature three pronounced peaks across the two time series: in early February, mid-March, and mid-April. These peaks correspond

Figure 4. (Color online) Announcement of the Sina Weibo Block



新浪微博公告

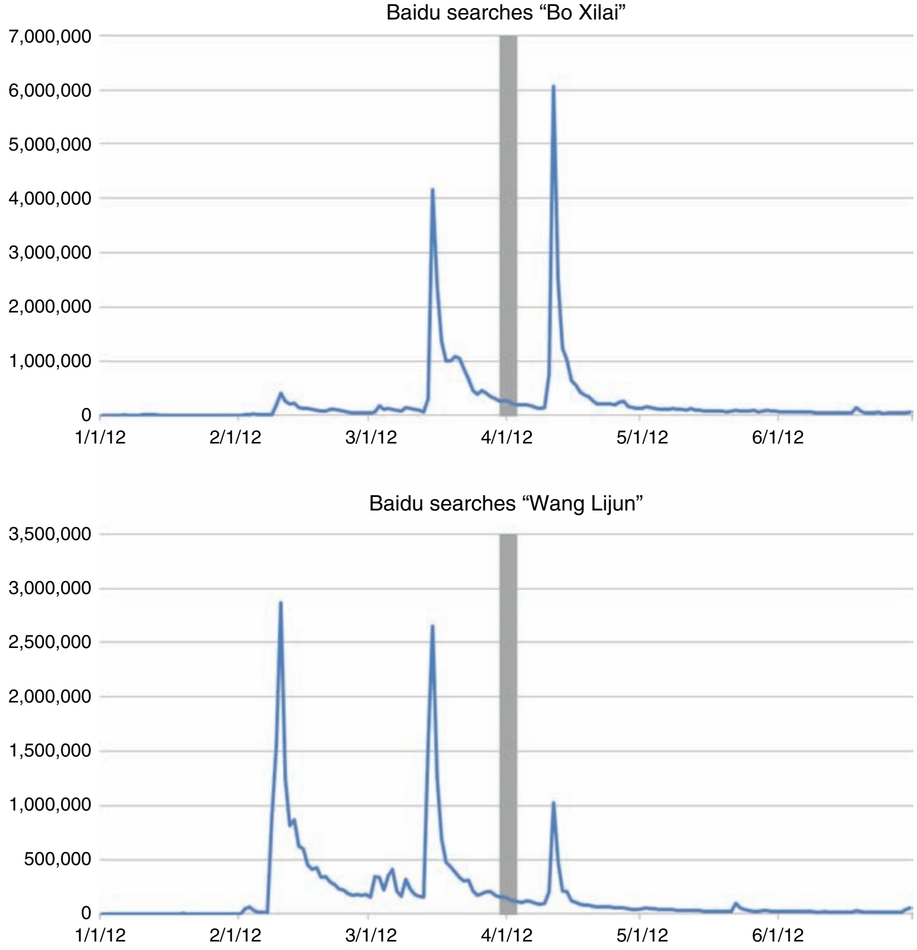
各位微博用户:

最近，微博客评论跟帖中出现较多谣言等违法有害信息。为进行集中清理，从3月31日上午8时至4月3日上午8时，暂停微博客评论功能。清理后，我们将再开放评论功能。进行必要的信息清理，是为了有利于为大家提供更好的交流环境，希望广大用户理解和谅解。感谢大家的支持。

新浪微博
2012年3月31日

Notes. See the main text for the English translation. The message was displayed on the Sina Weibo homepage (weibo.com) for the duration of the block.

Figure 5. (Color online) Time Series of the Daily Baidu Search Volume for the Names of the Two Politicians Involved in the Scandal



Note. The gray band indicates the three days of the block.

respectively to (1) Wang Lijun traveling to the U.S. consulate, (2) Bo Xilai's dismissal from his municipal post, and (3) Bo Xilai's suspension from the party's Central Committee. The censorship of Sina Weibo occurred between the latter two of the three peaks and during a time period in which searches for the two names were at a fairly low level. In other words, the block happened at a time when people paid little attention to the political scandal. We also analyzed similar graphs based on Google trends rather than the Baidu search data and found the time-series patterns to be very similar.

Second, we find no evidence that TV viewership changed as a function of the saliency of events related to the political scandal. We establish this pattern by using our later regression framework (see Section 4, Equation (1)) in which we regress episode-level (log) ratings onto show and weekday fixed effects and a dummy for the time period of the block. On top of these variables, we then also include the (log) search volume of both names in the regression. We find the coefficients on both variables are small and insignificant, whereas the coefficient on the block dummy remains significant, and its magnitude is not affected by the inclusion of the additional variables. We report results from this regression in Table A4 in the online appendix. We also show robustness to using the search-volume variables individually and jointly as controls as well as a specification with a series of dummy variables for the peaks of the scandal. Across all specifications, we find a clear and precise null effect of the search-volume and key-event dummy variables on TV show ratings.

Furthermore, to provide a more exhaustive assessment of interest in the scandal, we provide three additional pieces of evidence. First, we collect the Baidu search volume for a wider set of queries related to the scandal, namely, "Gu Kailai" (Bo Xilai's wife), "Bo Guagua" (Bo Xilai's son), "Chongqing" (the city where the scandal started), and "U.S. consulate" (Wang's visit to the U.S. consulate triggered the scandal). Second, we collect data on news coverage pertaining to the political events from Baidu news as well as Google news and compute the number of news articles published on either website that contain the name Bo Xilai or Wang Lijun. Third, Internet users in China were conceivably seeking out information about the scandal from sources of Western news or social media platforms that were blocked in mainland China (e.g., Twitter, Facebook, etc.). We capture some of this activity by collecting the Baidu search volume for the query “fan qiang,”[17] a keyword Chinese people use to find VPN software to circumvent the restrictive Internet control. Although the search for VPN software is not specific to the political scandal, we expect to see an increase in attempts to access censored sources during these politically sensitive events. Across all three sets of time series, (1) the larger set of Baidu queries, (2) news coverage, and (3) searches for the VPN software, we find no evidence for an increase in interest in the scandal around the time of the censorship, and these additional graphs show patterns similar to those in the graphs presented in Figure 5, with spikes around the three main events of the scandal. We report more details on how the graphs are computed in Section B of the online appendix.

4. Difference-in-Differences Analysis

Next, we analyze the magnitude of the ratings change during the censorship event and provide further evidence that the block affected TV viewership via its effect on the amount of TV-show-related microblogging activity on Sina Weibo. The latter is achieved via a set of DiD regressions and complements the analysis in Section 3 in terms of ruling out a direct effect of the block.

As a starting point for this analysis, we implement a “simple-difference” regression that analyzes the drop in ratings around the time of the block. Specifically, we run a regression of episode-level log ratings on a dummy for the three days of the block, show fixed effects, and day-of-the-week dummies

$$LogRating_{jt} = \alpha \cdot Block_t + \delta_j + Weekday'_t\gamma + \varepsilon_{jt}. \qquad (1)$$

The variable $LogRating_{jt}$ denotes the logarithm of the rating of show j on day t, $Block_t$ is a dummy variable equal to 1 for the three days of the block, δ_j is show j’s fixed effect, $Weekday_t$ is a vector of day-of-the-week dummies, and ε_{jt} is the regression error term. Standard errors are clustered at the show level. The results from this regression are reported in column (1) of Table 2 and show a significant drop in viewership during the Sina Weibo block.

This simple first regression might cause worry over two potential issues. First, the block (and the political events related to it) might have had a direct impact on viewership. Second, any unrelated event is potentially problematic if it happened during the same time window and also affected viewership. Based on the reasoning provided in Section 3, we believe a direct effect of the block is an unlikely scenario in our setting. Nevertheless, as an alternative way to deal with both issues just outlined, we implement an analysis that relies on the fact that different geographical areas, as well as different types of shows, should be differentially affected if the block affects viewership because of the change in microblogging activity, but not through any other channel.

Table 2. Difference-in-Differences Regressions: Geographical Differences

Dependent variable	(1) *Log rating*	(2) *Log rating*	(3) *Log rating*	(4) *Log rating*	(5) *Log rating*	(6) *Log rating*
Sample	Mainland China	HK and Mainland China	HK and Shenzhen (respective shows)	HK and Shenzhen (mainland shows)	24 cities in Mainl. China	24 cities in Mainl. China
Censor dummy	−0.017***	0.005	0.002	−0.008***	−0.010	−0.008
	(0.005)	(0.010)	(0.010)	(0.002)	(0.006)	(0.006)
Mainland × *Censor dummy*		−0.026**				
		(0.012)				
Shenzhen × *Censor dummy*			−0.035**	−0.017*		
			(0.014)	(0.010)		
Sina Weibo penetration × *Censor dummy*					−0.027*	
					(0.014)	
Above median penet. × *Censor dummy*						−0.016***
						(0.006)
Show FEs	Yes	Yes	Yes	Yes	Yes	Yes
Day of the week dummies	Yes	Yes	Yes	Yes	Yes	Yes
City FEs	n/a	n/a	n/a	n/a	Yes	Yes
Observations	7,899	11,427	11,427	15,798	189,576	189,576
Shows	193	325	325	193	193	193
R^2	0.881	0.964	0.951	0.774	0.479	0.479

Notes. The unit of observation is an episode in columns (1) to (4) and an episode/city combination in columns (5) and (6). Standard errors are clustered at the show level. HK, Hong Kong; FE, fixed effects.

***Significance at the 1% level; **significance at the 5% level; *significance at the 10% level.

4.1. Geography-Based Analysis

We start with an analysis along the geographical dimension, which leverages the fact that Sina Weibo is the predominant microblogging platform in mainland China but not in Hong Kong, where Twitter is also available. Because of this alternative, the use of Sina Weibo in Hong Kong is very low. Specifically, for the set of shows in our sample, Hong Kong users generate only 0.5% of the amount of tweets relative to the neighboring city Shenzhen in mainland China, which is comparable in population size. We take this finding as evidence that the usage of Sina Weibo is negligibly small in Hong Kong.

We run a DiD regression using data on all shows in mainland China as well as Hong Kong in which we interact the block dummy with a dummy for mainland China. We reiterate that the shows in Hong Kong and mainland China are different because Hong Kong consumers primarily watch local channels, which do not overlap with the major channels in mainland China. We are hence comparing different sets of shows between the two locations. Later, we run a further set of regressions based on a sample of overlapping shows. Similar to the previous regression, we include show and day-of-the-week dummies and cluster standard errors at the show level

$$\begin{aligned} LogRating_{jt} &= \alpha \cdot Block_t + \beta \cdot Block_t \cdot Mainland_j + \delta_j \\ &\quad + Weekday_t'\gamma + \varepsilon_{jt}. \end{aligned} \tag{2}$$

If the block affected TV viewership via lowering the amount of TV-show-related activity on Sina Weibo, we would expect the block to have affected viewership only in mainland China, and not in Hong Kong. Instead, if the block had a direct effect in the sense that the political events surrounding it led to lower TV viewing in general, we would expect to see the reduction in viewership in both Hong Kong and mainland China. Similarly, any unrelated event that happened during the period of the block would likely have affected both geographies comparably because of their cultural and economic similarity.

As evidence that Hong Kong constitutes a valid control group, we compute the correlations of the Baidu search volume regarding the scandal between mainland China and Hong Kong. The Baidu search-volume levels in Hong Kong and mainland China for the two involved politicians show strong correlations of 0.89 and 0.94, respectively. We also provide a plot of the time series of searches in Figure A6 in the online appendix, which shows that the volumes are highly correlated across the two locations with identical peaks in activity.[18] Hence, any direct effect of the political scandal that triggered the block is likely to have affected Hong Kong in a similar fashion as mainland China. The identifying assumption is hence that, absent the effect of the block on viewership through its impact on Sina Weibo, the time series of TV viewership in Hong Kong and mainland China would have been identical.

The results from this regression are reported in column (2) of Table 2. The primary coefficient of interest is the interaction of the block with the mainland dummy, which represents the DiD estimate. We find a negative and significant effect of −0.026. This effect is slightly larger, but not significantly different from, the simple-difference regression reported in column (1) of Table 2. We also note the coefficient on the censor dummy is small in magnitude and not significantly different from zero. In other words, for Hong Kong, we observe no difference in viewing behavior during the time period of the Sina Weibo block. Furthermore, in an effort to make the treatment and control groups even more similar to each other, we run the same regression as before but substitute the aggregated mainland data with city-level data from Shenzhen. Shenzhen neighbors Hong Kong, and the two cities are separated only by a river. The results from this regression, which are reported in column (3), are similar to the previous regression comparing Hong Kong and the entire mainland.

We then proceed to exploit one slightly different dimension of the data. The regressions in columns (2) and (3) both compare different sets of shows in Hong Kong and mainland China/Shenzhen. However, Hong Kong residents can receive mainland channels, and we are therefore able to compare the ratings for the *same set of shows* in Hong Kong and Shenzhen. To leverage the overlap in shows, we run the same specification as in column (3), but base the regression on the sample of mainland shows and their ratings in both Shenzhen and Hong Kong. Therefore, our sample for this regression includes every episode twice, once for Shenzhen and once for Hong Kong. Instead of show fixed effects, we include a separate fixed effect for each show/city pair. The results from this regression are reported in column (4), and, as expected, the drop in ratings is significantly larger in Shenzhen relative to Hong Kong. We also note the (uninteracted) censor dummy is negative and significant, most likely because mainland shows are tweeted about on Sina Weibo and hence consumers in Hong Kong can, in principle, access this information. Furthermore, the market share of mainland shows in Hong Kong is small, and therefore the selected group of Hong Kong–based consumers that watch mainland shows might also read the relevant comments on Sina Weibo.[19] Therefore, we would not necessarily expect to see no effect in Hong Kong. However, the stronger impact of the block in mainland China supports our assertion that the censorship affected mainland consumers more strongly because they rely relatively more on Sina Weibo.[20]

In a final set of regressions based on geographical differences in the strength of the effect, we explore heterogeneity across different cities in mainland China. In the same vein as the regression presented in the previous paragraph, these regressions compare the same set of shows across different geographical locations. Specifically, for the 24 mainland cities in our sample, we compute a proxy for the local usage of Sina Weibo. Similar to the comparison with Hong Kong, where Sina Weibo usage is close to zero, we would expect TV ratings in cities with higher penetration rates of Sina Weibo to be more affected by the block. To construct a proxy for the city-level penetration rates of Sina Weibo, we compute the number of Baidu searches for the terms "Sina Weibo Registration," "Sina Weibo Logon," and "Sina Weibo" relative to population size at the local level for six months before February 2012, when the political scandal started. To explore heterogeneity across cities, we run our baseline regression with an additional interaction term of the censor dummy and the local Sina Weibo usage variable[21] (as well as city fixed effects). Results from this regression are reported in column (5) of Table 2, and we find a negative effect on the interaction term that is significant at the 10% level. To interpret the magnitude of the coefficient, note that we scale the penetration-rate variable in such a way that it takes on values between 0 and 1. Therefore, the lowest penetration city experiences a rating decrease of 1%, whereas the highest penetration city sees ratings drop by 3.7%. We also implement a specification in which we interact the censor dummy with an indicator for whether the city has an above-median usage of Sina Weibo and again find a significant negative effect.[22]

4.2. Across-Show Analysis

In a second and complementary set of regressions, we analyze differences in the effect of the block across different shows in mainland China. Specifically, we group shows based on Sina Weibo activity outside of the block. If the block affected TV viewing only via its effect on Sina Weibo, we would expect to see a stronger decrease in ratings for shows with a more active presence on Sina Weibo and no effect for shows that have little to no Sina Weibo activity associated with them. We test this hypothesis using data on the average number of comments per episode of each show in the month before and after the block.[23] In other words, we categorize shows by the "normal" amount of commenting activity for an episode that did not air during the block.

In a first regression, we categorize shows into three equally sized bins according to their level of activity on Sina Weibo. We note the lowest activity category contains shows with less than one comment per episode; hence, these shows have almost no activity. Medium- and high-activity shows are characterized by at least 3 and 45 comments per episode, respectively. Column (1) of Table 3 shows results from a regression in which we interact the block dummy with dummies for the different activity levels. As before, we control for show fixed effects and day-of-the-week dummies, and cluster standard errors at the show level

$$\begin{aligned} LogRating_{jt} = \alpha \cdot Block_t + \alpha_M \cdot Block_t \cdot Medium_j \\ + \alpha_H \cdot Block_t \cdot High_j + \delta_j \\ + Weekday'_t \gamma + \varepsilon_{jt}. \end{aligned} \tag{3}$$

The variables $Medium_j$ and $High_j$ denote dummies for two of the activity groups. Low-activity shows constitute the omitted category, and the block dummy captures the effect on these shows. We find the effect for the low-activity shows is small in magnitude and insignificant. The medium-level shows experience a stronger drop of 1.3% in their ratings (i.e., 0.5% + 0.8%). However, the decrease for medium-activity shows is not significantly different from the effect for low-activity shows. For shows with a strong presence on Sina Weibo, we find an effect of larger magnitude. For this category of shows, ratings dropped by 3.1% during the block (i.e., 0.5% + 2.6%), and the estimated coefficient is significantly different from the effect for low-activity shows.[24]

To probe the robustness of the results, we also interact the average number of comments per episode (in units of 100,000 comments) linearly with a dummy for the time period of the block. We find a significant effect for the interaction term, which we report in column (2) of Table 3. Because the distribution of show-related tweets is highly skewed, we prefer the first specification, which splits shows into three bins rather than the interaction with a linear term.

Apart from differences in activity levels on Sina Weibo, we can also characterize shows along other dimensions that might have influenced how strongly they were affected by the censorship event. We thus proceed to analyze the differential impact of the block on different show genres, new versus old shows, and so on. Our analysis in this regard is based on the different show attributes that were presented in the lower panel of Table 1. We first analyze differences in the impact of the block for shows belonging to different genres. Specifically, we can categorize shows in our sample into four major categories: TV series (23 shows), reality shows (40), news/reporting on current events (50), and children's shows (16). The remainder of the shows (64) belong to a variety of genres and are treated as a residual category. Results from a regression that interacts a dummy variable for each of the five categories with the censor dummy is reported in column (3) of Table 3.[25] In terms of magnitude, we observe large negative coefficients for TV series, reality shows, and news shows, but only small coefficients for children's shows and

Table 3. Difference-in-Differences Regressions: Across-Show Differences

	(1)	(2)	(3)	(4)	(5)
Dependent variable	*Log rating*	*Log rating*	*Log rating*	*Log rating*	*Log rating*
Censor dummy	−0.005	−0.012**			
	(0.005)	(0.005)			
Medium Weibo activity ×	−0.008			−0.007	
Censor dummy	(0.011)			(0.009)	
High Weibo activity ×	−0.026**			−0.024**	
Censor dummy	(0.012)			(0.011)	
Weibo activity (unit: 100,000 com.) ×		−0.029**			−0.030***
Censor dummy		(0.012)			(0.011)
TV series ×			−0.025	−0.005	−0.015
Censor dummy			(0.020)	(0.021)	(0.016)
Children's show ×			−0.003	0.002	−0.003
Censor dummy			(0.007)	(0.007)	(0.007)
Current events show ×			−0.024***	−0.015**	−0.021***
Censor dummy			(0.006)	(0.006)	(0.006)
Reality show ×			−0.020	−0.005	−0.002
Censor dummy			(0.015)	(0.014)	(0.015)
Other shows ×			−0.008	−0.001	−0.008
Censor dummy			(0.007)	(0.009)	(0.007)
Show FEs	Yes	Yes	Yes	Yes	Yes
Day of the week dummies	Yes	Yes	Yes	Yes	Yes
Observations	7,899	7,899	7,899	7,899	7,899
Shows	193	193	193	193	193
R^2	0.881	0.882	0.881	0.881	0.882

Notes. The unit of observation is an episode. Standard errors are clustered at the show level. com., comments; FE, fixed effects.

***Significance at the 1% level; **significance at the 5% level; *significance at the 10% level.

the residual category. Although the five coefficients are not significantly different from each other, we do find that TV series, reality shows, and news shows as a group experience a significantly larger decrease than the other two genres.[26]

Furthermore, the genres with larger effect sizes are tightly correlated with the types of genres that experience higher Sina Weibo activity. The data on activity levels across genres were presented in Table 1, and we find average numbers of comments per episode of 24,504, 48,631, and 5,358 for TV series, reality shows, and news shows, and much fewer comments, 26 and 48 per episode, for children's shows and other shows, respectively. We therefore conjecture that the differential impact we observe across different show genres proxies for the key driver of effect heterogeneity, which is the average activity level on Sina Weibo of each show. To further explore this assertion, we estimate a regression that includes interactions of the censor dummy with both show genres as well as with activity levels (see column (4)). When doing so, the effect sizes of the genre interaction terms decrease substantially, whereas the effect magnitudes on the interaction terms with activity levels are very similar to those for the specification in column (1), which includes only those terms. The one exception is the effect on news shows, which is significantly different from zero; however, the five genre interaction terms are not significantly different from each other. The results are similar when including genre dummy interaction together with a linear interaction term with Weibo activity in column (5). We hence conclude that genre has no major role in predicting the ratings drop during the block, after controlling for differences in the shows' average activity level.

Finally, we also explore heterogeneity between shows that started airing recently versus longer-running shows, shows that air daily versus at a lower frequency, and reruns versus new shows. When exploring effect heterogeneity along each of these three dimensions, we find no significant differences across shows aired at different frequencies or for new versus established shows.[27] The absence of a significant effect is most likely due to the fact that the activity level in terms of the average number of comments does not differ much across those two dimensions. The only characteristic that is predictive of the impact of the block is whether the show is a rerun or a current show. Reruns are significantly less affected by the censorship. Similar to the heterogeneity patterns across genres, this finding is most likely driven by the fact that reruns do receive less attention on Sina Weibo, and the average number of comments is much smaller (59 per episode) for reruns relative to current shows (16,000 comments per episode).[28]

4.3. Robustness Checks

One concern with any DiD setting is the existence of preexisting differential time trends between the treatment and control groups. This concern typically applies to regressions that compare treatment- and control-group observations before and after a policy change. The identifying assumption in such a setting is that the treatment group, had it not been treated, would have followed the same trajectory over time as the control group. Therefore, the existence of differential time trends might cast doubt on the validity of this assumption.

In our context, this issue is less likely to be a concern. First, we study the evolution of TV show ratings over a relatively short time horizon of two months; hence, strong time trends in TV viewership are unlikely. Second, our setting is slightly richer than other DiD settings, because the treatment is temporary. We thus have observations for both before and after the treatment period. Therefore, the identification argument boils down to the assumptions that during the three days of the block, the treatment group would have evolved in the same way as the control group had it not been treated. Differential time trends between both groups are therefore less likely to lead to a spurious result.

We also explicitly test for the presence of differential time trends and assess whether our key coefficients of interest are robust to including them. We implement this robustness check for the main regressions presented in this section. Specifically, we rerun four DiD regressions across geographies and across types of shows (one regression of Hong Kong versus mainland China, two regressions across cities within mainland China, and one regression across shows with different Weibo activity levels), but also include a linear time trend and a linear time trend interacted with a treatment-group dummy for each respective regression. We report the results from this set of regressions in Table A5 in the online appendix. As expected, we find little evidence of any time trends for either the treatment or control group in any of the specifications. Furthermore, the coefficients of interest of the censor dummy interacted with the relevant treatment-group dummy are almost unchanged across all specifications.

As a further test, we also implement a set of placebo regressions in which we move the treatment period from late March/early April to either mid-March or mid-April. To make things as comparable as possible, we implement two placebo treatment periods that are three days long, which is equal to the duration of the actual block. We run two placebo tests for each of the four specifications listed above, and out of 10 relevant coefficients across all regressions, we find only 2 to be significant (one at the 1% level, the other at the 10% level).[29] In both cases, the sign of the estimated coefficient has the opposite sign relative to the “true” treatment coefficient. Furthermore, we run the identical set of falsification tests for a one-week shift of the block. This approach allows us to hold fixed the days of the week during which the (placebo) block happens. We implement the analysis for the week before and the week after the block across all four specifications and find 1 coefficient out of 10 is significant at the 10% level. All other coefficients are statistically insignificant. The results from these placebo tests provide further evidence that differential time trends are unlikely to contaminate our results.

Finally, we run a robustness check with regard to the computation of standard errors, which are clustered at the show level in all of the regressions reported above. As an alternative, we follow Bertrand et al. (2004) and compute the average rating of each show separately for all episodes during and outside of the block, thus reducing the data to exactly two observations per show (one for during the block and one for outside the block).[30] For each pair of treatment and control groups across three specifications,[31] we then compute the DiD estimate and the corresponding standard error. We find our results remain qualitatively similar. In the case of the comparison of shows in mainland China and Hong Kong, as well as shows with different microblogging activity levels, precision increases when collapsing the data. For the specification based on cities with different Sina Weibo penetration rates, precision decreases and the estimated effect is significant only at the 10% level.[32]

4.4. Firms’ Response and Other Marketing Activity

One further concern with any external shock to one particular kind of marketing activity is that firms might react by adjusting other types of marketing activity. Because of the short duration of the block, we think such a scenario is unlikely in our case. Furthermore, the censorship was not announced beforehand, and firms were unlikely to have been able to anticipate the block. Therefore, the short duration and suddenness of the block constitutes a feature of our natural experiment that provides a clean setting to isolate the effect of changing only one type of marketing activity (WOM in this case). Any longer-term shock is likely to trigger a response from firms, which would constitute a further obstacle to correctly identifying the effect of interest.

5. Elasticity Estimate: Instrumental Variable Regression

In this section, we proceed to measure the impact of comments on ratings, which is the key estimate of interest to evaluate the impact of online WOM in our context. So far, we have analyzed the impact of the block on ratings. We now add data on microblogging activity at the episode level and estimate the impact of comments on ratings by isolating the variation in comments caused by the censorship event. Specifically, we run an IV regression in which we regress ratings on the total number of comments related to the specific episode and instrument the number of comments with a dummy for whether the episode aired during the block. The DiD regressions in Section 4 and the time line of the scandal presented in Section 3 both provide evidence that the block affected TV show ratings via its effect on Sina Weibo activity and did not directly affect ratings. The absence of any direct effect of the block on ratings is the key identifying assumption and allows us to exclude the block from the second stage of the IV regression.

Similar to the regression framework used in Section 4, we control for show fixed effects and day of the week, and cluster standard errors at the show level. Formally, we run the following regression:

$$LogRating_{jt} = \alpha \cdot LogComments_{jt} + \delta_j + Weekday'_t \gamma + \varepsilon_{jt}, \qquad (4)$$

where $LogComments_{jt}$ denotes the (log) number of comments related to an episode of show j, which aired on day t. This variable is instrumented with a dummy for whether Sina Weibo was blocked on day t.

5.1. First Stage

Before proceeding to the actual first stage of our IV regression, we assess the effect of the block on activity on Sina Weibo more broadly. To this end, we analyze the time series of show-specific activity in the simplest possible way by calculating the number of comments for each day in March and April 2012 for each show contained in our sample (regardless of whether an episode actually aired on the specific day). We then regress the number of comments onto show fixed effects and weekday dummies, as well as a dummy for the three days of the block.[33] Columns (1) and (2) in Table 4 report the results using the number of comments in either levels or logs. For both specifications, we find the block caused a substantial drop in the number of comments.[34]

We repeat the same type of regression for the other three types of user involvement on Sina Weibo: tweets, retweets, and likes. We report the results from these regressions in Table A6 in the online appendix. With the exception of a significant positive effect (only in the log specification) for retweets, we find no evidence that the block affected any type of activity on Sina Weibo other than comments. The fact that other types of activity are unchanged during the duration of the censorship is important because we are attributing the change in ratings to the change in comments during the block. If other types of activity also changed, the exclusion restriction would be violated because the block would have affected ratings not only via comments but also through its indirect impact on other types of microblogging activity.[35] We also note that ideally we would like to study the impact of all four activities on ratings, but such an analysis would require four separate instruments, whereas we have access only to one.

We next present the actual first stage of our IV specification, which we run at the episode level. To implement such a regression, we need to associate Sina Weibo activity over a specific time period with each episode for which we have ratings data. We define the number of relevant comments pertaining to a specific episode as the number of comments that mentioned the show on the same calendar day on which the particular episode aired. The results from these "episode-centric" regressions are reported in columns (3) and (4) of Table 4 and are very similar to their counterparts using daily data in the first two columns. We also note that, especially for the log specification, the first stage estimate is highly significant, as the F-statistic on the censor dummy (the excluded instrument) shows, which mitigates weak-instrument concerns (Rossi 2014).

Preshow and anticipated postshow comments can conceivably influence ratings. Both are affected by the block and the total episode-level comments measure comprises both types of comments. Our IV regression hence does not disentangle these possible channels and results are consistent with either preshow or anticipated postshow comments causally affecting ratings (or both). We explore this distinction in detail in Section 6.

5.2. Second Stage and Effect Magnitude

We report results from the second stage in columns (5) and (6) of Table 4 for the level and log specifications, respectively. In our preferred log–log specification, we find a statistically significant coefficient of 0.016, and hence doubling the number of comments leads to a 1.6% increase in ratings.[36] To gauge the magnitude of this effect, we first note that doubling the number of comments leads to a movement of roughly 20% of a standard deviation of the typical within-show fluctuation in ratings over time (see Table 1) and thus constitutes an economically important effect.[37]

At the same time, our estimate is substantially lower than most estimates in the previous literature. For instance, Sonnier et al. (2011) estimate a short-run elasticity of (positive) WOM of 0.64, Dhar and Chang (2015) find an elasticity of 1.04 for music sales, and Liu (2015) estimates an elasticity of 0.59 for box office revenue in the release week of a movie. These estimates (as well as those from several other papers in the WOM literature) are more than an order of magnitude larger than our

Table 4. Instrumental Variable Regressions

Dependent variable (DV)	(1) *Number of daily comments*	(2) *Log number of daily comments*	(3) *Number of episode-level comments*	(4) *Log number of episode-level comments*	(5) *Rating*	(6) *Log rating*
Type of regression	OLS	OLS	IV 1st stage	IV 1st stage	IV 2nd stage	IV 2nd stage
Standard deviation of DV (control for show FEs)	47,503	1.304	33,679	1.463	0.158	0.090
Number of episode-level comments (unit: 10,000)					0.024*** (0.009)	
Log number of episode-level comments						0.016*** (0.005)
Censor dummy	−10,351** (5,172)	−1.003*** (0.086)	−14,751** (6,504)	−1.066*** (0.104)		
F-stat. on censor dummy	4.00	136.38	5.14	105.14		
Show FEs	Yes	Yes	Yes	Yes	Yes	Yes
Day of the week dummies	Yes	Yes	Yes	Yes	Yes	Yes
Observations	10,126	10,126	7,899	7,899	7,899	7,899
Shows	166	166	193	193	193	193
R^2	0.483	0.789	0.595	0.839	0.845	0.877

Notes. The unit of observation is a show/day combination in columns (1) and (2) and an episode in the remaining columns. Standard errors are clustered at the show level in all regressions. The regressions in columns (1) and (2) also include a dummy for whether an episode of the show was aired on the specific day. OLS, Ordinary least squares; FE, fixed effects.

***Significance at the 1% level; **significance at the 5% level; *significance at the 10% level.

elasticity estimate of 0.016. Although these papers do study WOM in different product markets and on different WOM platforms, we suspect that part of the discrepancy is due to the fact that previous papers were limited in their ability to deal with issues of endogeneity because of the nature of their data. We note that one other more moderate elasticity estimate is from Lovett and Staelin (2016), who find an elasticity of 0.04.

More important for the optimal allocation of marketing budgets across channels, we obtain an estimate for the impact of WOM that is smaller than the magnitude levels found for TV advertising of around 0.03 (Gordon and Hartmann 2013, Tuchman 2016, Shapiro 2016). Therefore, the seemingly higher impact of WOM relative to traditional advertising when comparing correlational studies might be an artifact of endogeneity bias rather than actual differences in effectiveness. Of course, we estimate the impact of WOM only for one particular product market and one specific type of WOM. Hence, whether the more modest effect size constitutes a finding that holds across a variety of markets and forms of WOM is an issue that is beyond the scope of this study and is left for future research.

Finally, it is instructive to assess, using a back-of-the-envelope calculation, how many households changed their viewership behavior relative to the number of active and passive Sina Weibo users. The average show in our sample has a rating of 0.434% (i.e., 0.434% of households were watching the show). In 2012, China had 456 million households, and the block moves ratings by 1.6%. Therefore, 450,000,000 × 0.00434 × 0.016 = 32,000 fewer households watched the average episode. With respect to the amount of microblogging, the average episode received 11,311 comments. Under the assumption that these comments originate from different users, and assuming a ratio of readers to contributors of 100:1,[38] the block made 3% of Sina Weibo readers (32,000/(11,300 × 100)) not tune into the respective show. We note this calculation depends crucially on the ratio of readers to contributors, which we do not observe in the data. Hence, we use an external number as an approximation.

5.3. The Nature of the Experimental Variation and Its Impact on Effect Magnitude

As with any natural experiment, our analysis is predicated on the specific variation induced by the natural experiment. Several aspects of this variation are useful to keep in mind when interpreting the magnitude of the effect and its external validity.

First, our natural experiment provides a substantial shift in microblogging activity by fully eliminating it during the block. This type of large deviation is likely to yield a different effect on ratings than a change in microblogging activity at the margin. Given the nature of our shock, we are unable to trace out the response curve in our setting. However, we posit that the most likely scenario is one with decreasing returns from additional microblogging activity, and hence the marginal effect will be smaller than the one estimated from a larger deviation from the status quo level.

Second, the censorship event eliminated microblogging activity for all shows rather than for any one individual show. Therefore, the effect on ratings from a unilateral change, holding microblogging for all competitor shows constant, could conceivably lead to a larger reaction in terms of ratings, because consumers might substitute to other shows that are still receiving comments. To assess whether such competitive effects are likely to matter in our context, we investigate whether consumers substitute primarily between shows or toward the outside option during the block. We find the latter to be the dominant channel, and the primary source of the loss in ratings is due to consumers watching less TV overall. Therefore, the effect of a unilateral decrease in microblogging might not be very different from the effect due to a decrease across all shows. We describe the analysis of substitution patterns in detail in Section E of the online appendix.

Finally, our estimate isolates the elasticity of comments while holding all other types of microblogging activity constant. One would expect a larger elasticity when increasing all activity by a certain percentage, rather than just increasing comments.

6. Mechanism

We now turn to analyzing the behavioral mechanism underlying the estimated effect of microblogging on demand. To structure our analysis of the mechanism, we base it on the three main conceptual frameworks with regard to the impact of advertising outlined in Bagwell (2007): informative and persuasive effects and complementarity between advertising and product consumption. Although we study WOM rather than traditional advertising, the distinction between the three possible channels applies here. Informative effects can arise if microblogging reminds the consumer of a show's existence or provides additional information about features of the show that the consumer values. Persuasive effects increase the consumers' appreciation of the show without delivering information about the show or its content. Finally, complementarity between WOM and TV show viewership arises if consumers derive a higher utility from viewing the TV show when they are also able to engage in microblogging (either actively or passively).

Among the three mechanisms, only the complementarity channel allows for an active role of the consumer in her decision to obtain information. Informative and persuasive effects are entirely passive; that is, the consumer gets exposed to certain information by reading content on Sina Weibo, and this exposure might affect her choice regarding which TV show to watch. Instead, complementarity implies the consumer actively decides how much information to consume. In the context of advertising, the consumer might decide to watch certain ads as a function of her past consumption of the product (Becker and Murphy 1993). In the case of microblogging, the case for an active role is even stronger because reading tweets and comments is arguably a more conscious choice than watching an ad. Furthermore, the consumer might also decide to contribute information by commenting herself, a possibility that is absent with regard to firms' advertising. We therefore believe studying complementarity is particularly interesting in this context.

First, we separate complementarity from informative/persuasive effects based on the timing of comments. Second, we distinguish between informative and persuasive effects based on the actual content of comments. The latter part requires additional data on microblogging content. We gathered such data for a random sample of around 12,000 comments[39] across all shows in our sample, and we assign each comment to one of the following categories: informative, expressing sentiment, or neither. In the case of sentiment, we further categorized these comments into positive or negative sentiment. Across all shows in our data, we find that 25% of comments express sentiment, of which 24% are positive and 1% are negative. Only 1% of comments contain informative content (e.g., the time at which the episode airs). In Table A7 in the online appendix, we provide more detailed descriptive statistics on the distribution of content attributes across shows.

6.1. Consumption Complementarity

We first turn to disentangling the complementary function of microblogging from the informative and persuasive channels. For comments to affect consumers via the latter two channels, consumers need to be exposed to comments on Sina Weibo before the show actually airs. Such preshow comments can inform consumers about the show or persuade them to watch the show due to the enthusiastic comments of other users. Therefore, when preshow comments disappear during the block, some subset of consumers will not be informed about or persuaded to watch the show, which might explain the decrease in ratings.

Alternatively, consumers might gain utility from engaging in microblogging (either reading or contributing) *after* the show to discuss their opinions about the episode. In this scenario, when the censorship event removes the ability to comment after a show has aired, the consumer's utility from watching the show decreases, and hence fewer consumers tune in. In principle, complementarity could also apply to preshow comments. For instance, consumers might have a discussion about the plot of the upcoming episode, and being able to engage in such a discussion increases the utility from watching the show. Therefore, any impact of postshow comments is consistent only with the complementarity channel, whereas an impact of preshow comments is consistent with all three channels and will require further investigation. We also note that for postshow comments to influence viewership, we require consumers to be able to anticipate the (lack of) consumption of postshow microblogging activity when deciding which show to watch. Because of the salient announcement of the block (see Figure 4), this condition is likely to be met in our context.

Based on this reasoning, the ideal variation to distinguish between complementarity and the other channels would be random variation in whether pre- or postshow comments were disabled across shows. In our case, the censorship event eliminated both pre- and postshow comments for all shows in our sample, and we hence do not have a source of exogenous variation that differentially affects both types of comments. In the absence of such variation, we turn to exploring effect heterogeneity across different types of shows in the following way: for each show in our data, we compute the average number of comments per episode that were posted before and after the show aired.[40] We then analyze whether the block differently affected shows that typically had more activity before versus after the show. The idea behind this analysis is that shows with predominantly preshow activity suffered mostly in terms of losing that preshow activity during the block, and vice versa for shows with mostly postshow comments. Hence, if postshow activity is the main driver of ratings, we should see shows with mostly postshow activity experience a larger ratings drop during the block.

We report results from the relevant regressions in Table 5. We start by replicating column (1) of Table 3 in the first column, which splits shows by the average amount of activity (regardless of the timing of comments) and interacts dummies for high- and medium-activity shows with a dummy for the block. We then run the same regression, but separately split the numbers of pre- and postshow comments into three equally sized bins each and include interaction terms of the block dummy with those variables in the regression reported in column (2). Doing so, we find that shows with high postshow microblogging activity experi-

Table 5. Timing and Content: The Differential Impact of Weibo Activity

	(1)	(2)	(3)	(4)
Dependent variable	*Log rating*	*Log rating*	*Log rating*	*Log rating*
Censor dummy	−0.005	−0.001	−0.002	−0.002
	(0.005)	(0.007)	(0.007)	(0.007)
Medium daily activity × *Censor dummy*	−0.008			
	(0.011)			
High daily activity × *Censor dummy*	−0.026**			
	(0.012)			
Medium preshow activity × *Censor dummy*		−0.007	−0.007	−0.007
		(0.010)	(0.012)	(0.012)
High preshow activity × *Censor dummy*		0.011	0.024	0.028
		(0.020)	(0.019)	(0.019)
Medium postshow activity × *Censor dummy*		−0.007	−0.007	−0.008
		(0.009)	(0.009)	(0.009)
High postshow activity × *Censor dummy*		−0.041**	0.001	0.005
		(0.020)	(0.018)	(0.019)
Medium postshow (any) sentiment comments × *Censor dummy*			0.007	
			(0.014)	
High postshow (any) sentiment comments × *Censor dummy*			−0.060***	
			(0.016)	
Medium postshow positive sentiment comments × *Censor dummy*				0.017
				(0.014)
High postshow positive sentiment comments × *Censor dummy*				−0.039**
				(0.017)
Medium postshow negative sentiment comments × *Censor dummy*				−0.017
				(0.014)
High postshow negative sentiment comments × *Censor dummy*				−0.041**
				(0.018)
Show FEs	Yes	Yes	Yes	Yes
Day of the week dummies	Yes	Yes	Yes	Yes
Observations	7,899	7,899	7,899	7,899
Shows	193	193	193	193
R^2	0.881	0.881	0.881	0.881

Notes. The unit of observation is an episode. Standard errors are clustered at the show level. FE, Fixed effects.

***Significance at the 1% level; **significance at the 5% level; *significance at the 10% level.

enced the strongest decline in ratings during the censorship event. Instead, shows with high preshow activity did not experience a drop in ratings during the block of Sina Weibo. Based on the reasoning presented above, we take these patterns as evidence of a complementarity effect rather than an informative or persuasive effect.

We then delve deeper into the nature of postshow comments that are predictive of the impact of the block. We first include the number of postshow comments expressing sentiment (again using a three-bin split) on top of the total number of postshow comments. Doing so, we find the coefficient on high total postshow comments becomes insignificant, whereas the coefficient on high postshow comments expressing sentiment is negative and significant (see column (3) of Table 5). These results show that enthusiastic post-show discussions are the primary driver of complementarity with show viewership. The post-show comments that do not express sentiment do not seem to affect viewership of the show.

When we further split comments into those containing positive and negative sentiments in column (4) of Table 3, we find both types of comments lead to a larger drop in ratings during the block. This finding is particularly interesting because most prior papers that study valence find negative WOM leads to a decrease in product demand (see, e.g., Sonnier et al. 2011). However, most of these papers have a mechanism in mind whereby WOM affects the consumer prior to making a purchase decision, and hence negative content will lead to a decrease in purchases. Instead, in the case of complementarity, negative comments are not necessarily a bad thing. Instead, they could simply be the sign of a heated debate among viewers of the show. We conclude that a stronger emotional engagement on Sina Weibo after an episode airs is the type of activity that most strongly affects viewership.

We also note the webpage layout for comments on Sina Weibo is one that allows for discussion between users, because it displays comments in a common thread underneath each original tweet (see Figure 3). The users who contribute comments do not need to follow each other and can use the comment section to participate in a conversation with others who are interested in the same subject. The interactive element of comments is therefore conducive to generating complementarity between TV show viewership and postshow consumption of microblogging content. Furthermore, we conjecture this mechanism might be unique to comments and applies less to tweets and retweets, because they do not allow for a dialogue between users in the same way comments do. Therefore, our results with regard to consumption complementarity might not extend to other forms of microblogging activity. As an additional piece of evidence that user engagement is high for postshow comments, we also note that the fraction of comments expressing sentiment is higher for postshow relative to preshow comments (see Table A7 in the online appendix).

Finally, we investigate whether the level of commenting activity either before or after a show is systematically correlated with show characteristics such as genre, whether the show is a rerun, and so on. We find that shows with high preshow commenting activity and shows with high postshow commenting activity have similar characteristics (i.e., they tend to be of a similar genre, are less likely to be reruns, etc.).[41] This finding reassures us a correlation of high postshow (relative to preshow) activity with other show characteristics is not the reason for a differential decrease in ratings for shows with a high level of postshow comments. Table A8 in the online appendix provides detailed descriptive statistics on the distribution of show characteristics as a function of commenting activity before and after a particular show.

6.2. Informative and Persuasive Effects

Because of the support we find for complementarity over an informative or persuasive effect, the distinction between the latter two is more secondary in our context. Nevertheless, the insignificant impact of preshow comments could be masking the fact that some types of preshow comments do in fact influence ratings. We therefore run a set of regressions (not reported) in which we include the number of preshow comments that are informative or express sentiment. The idea behind this regression is that an informative effect would predict that comments that are informative in nature have a larger impact, whereas a persuasive effect would suggest comments expressing sentiment have a greater impact. When including both types of comments both individually and separately in the regression, we find they are consistently insignificant. We therefore conclude that our analysis does not support the notion that WOM affects ratings via an informative or persuasive channel.

6.3. Other Patterns in Support of Consumption Complementarity

A few other patterns of effect heterogeneity across different types of shows also provide some additional evidence for the consumption complementarity channel identified above. First, we find no difference in the ratings drop during the block for new versus established shows (see Endnote 27). If informative effects were important, we would expect to see a larger impact for new shows of which consumers are less aware. The absence of a differential effect for new shows is therefore consistent with the absence of informative effects.

Second, we present evidence in Section 7 that WOM does not have any longer-term effects on ratings. In other words, after controlling for comments pertaining to the current episode, comments that were posted about prior episodes do not have any impact on current ratings. If the anticipation of the microblogging activity after the show is the main driver of ratings, we would not expect past microblogging activity to affect current ratings. Instead, other types of mechanisms, most importantly the persuasive channel, are more likely to imply dynamic effects in the sense that consumers who are exposed to WOM in multiple time periods are more likely to be persuaded to watch a specific show. We therefore see the absence of dynamic effects as an additional piece of evidence in support of the complementarity channel.

7. Dynamics and Long-Term Effects

In this section, we explore whether changes in microblogging activity have any longer-term impact beyond influencing ratings of the current episode. To analyze dynamic effects of microblogging on TV show ratings, we rerun our baseline regression and include lagged comments (with different lag structures) as additional regressors. In keeping with our identification strategy for the contemporaneous effect of comments on ratings, we instrument each lagged-comments variable with an equivalently lagged censor dummy variable. In this way, we are exploiting only variation in past commenting activity that is caused by the censorship event.

Column (1) in Table 6 replicates our baseline regression without lagged effects. In column (2), we add comments pertaining to the previous episode of the same show as an additional regressor. As before, the relevant set of comments is defined as all comments posted on the day the specific episode aired. Lagged comments are defined in the same way, but in relation to the day on which the previous episode aired. When including the "one-episode-lagged" number of comments, we find the coefficient on the lagged term to be small and statistically insignificant. The coefficient on contemporaneous comments remains almost unchanged and statistically significant. We further explore regressions with up to three periods of lags, and consistently find small and insignificant effects on all lagged terms. We also note the first-stage regressions have strong predictive power. Table 6 reports the F-statistic for all first-stage regressions for each specification.[42]

In summary, the results reported above show no evidence of dynamic WOM effects in our setting, and only contemporaneous microblogging activity has an impact on TV show viewership. This finding is in contrast to several papers in the previous WOM literature that have emphasized the importance of dynamic considerations (Villanueva et al. 2008, Sonnier et al. 2011). At the same time, as discussed in Section 6, the absence of a dynamic effect is consistent with Sina Weibo comments affecting ratings because of consumption complementarity between TV viewing and microblogging.

8. Conclusion

Promoting products through social media websites such as Twitter, and allowing users to discuss and voice their opinions about a product, has become a common practice for many firms. Yet whether this new market-

Table 6. Dynamic Effects

	(1)	(2)	(3)	(4)
Dependent variable	*Log ratings*	*Log ratings*	*Log ratings*	*Log ratings*
Type of regression	IV 2nd stage	IV 2nd stage	IV 2nd stage	IV 2nd stage
Log number of comments	0.016*** (0.005)	0.015*** (0.005)	0.015*** (0.004)	0.017*** (0.005)
Lagged log comments ($n-1$)		0.00005 (0.00496)	−0.00437 (0.00516)	−0.00398 (0.00514)
Lagged log comments ($n-2$)			0.00530 (0.00512)	0.00396 (0.00568)
Lagged log comments ($n-3$)				0.00048 (0.00487)
First-stage F-stat (s)	105.14	59.54, 64.31	40.54, 47.50 43.75	30.72, 36.13 38.07, 32.13
Show FEs	Yes	Yes	Yes	Yes
Day of the week dummies	Yes	Yes	Yes	Yes
Observations	7,899	7,733	7,567	7,399
Shows	193	193	193	193
R^2	0.877	0.880	0.880	0.882

Notes. The unit of observation is an episode. Standard errors are clustered at the show level. Lagged comments are instruments with equivalently lagged censor dummies. FE, Fixed effects.

***Significance at the 1% level; **significance at the 5% level; *significance at the 10% level.

ing channel can effectively enhance demand remains unclear. Similar to evaluating the effectiveness of other types of marketing activity, measuring the effect of microblogging using field data is challenging, and the correlation between microblogging activity for a product and its demand does not necessarily imply causality. Obtaining credible evidence on the impact of WOM is particularly challenging because the firm does not directly control WOM, and hence field experiments are difficult to implement. In this paper, we leverage a natural experiment to identify the causal effect of WOM on product demand and investigate the mechanism through which WOM affects demand.

Several novel findings emerge from our analysis. First, the magnitude of the estimated WOM elasticity is significantly lower than the magnitudes obtained in previous studies. Our findings therefore caution against overstating the impact of WOM, and highlight the importance of using exogenous variation to obtain a causal estimate. The difference in magnitude is particularly striking in our setting: our estimate is more than an order of magnitude lower than many other estimates in the WOM literature. Furthermore, although correlational estimates suggest WOM is more effective than TV advertising (0.12 for TV advertising; Sethuraman et al. 2011, versus 0.2 for WOM; You et al. 2015), the opposite is true for causal estimates in both realms (0.016 for WOM obtained in this study versus 0.03 for TV advertising). We note that the set of causal estimates for both WOM and TV advertising is much smaller, and therefore further work is required to establish whether the relative performance of both channels generalizes beyond the existing studies.

Second, we analyze the mechanism by which WOM influences demand in our setting, and find that complementarity between TV and microblogging consumption is the main driver. Particularly, we find the anticipation of more postshow microblogging activity leads to an increase in TV ratings. Several intriguing managerial implications emerge from this analysis. First, rather than attempting to increase WOM prior to the show airing to remind or persuade consumers, fostering an active discussion after the show is more important. Second, contrary to other mechanisms, consumption complementarity does not entail any dynamic effects of microblogging, which is consistent with what we find empirically. Finally, our findings show both positive and negative postshow comments have a positive impact on ratings, and therefore a discussion that engages consumers appears to be the key driver of ratings. Our findings with regard to negative comments are particularly interesting because one would expect them to decrease ratings if the effect operated through informative or persuasive channels. Instead, the case is less clear for our proposed channel of complementarity with postshow microblogging, and the conventional wisdom of avoiding negative content might not necessarily apply here.

In summary, both with regard to the effect magnitude as well as the proposed mechanism and its implications, our paper provides a set of novel findings that further the understanding of the impact of WOM on demand.

Acknowledgments

The authors thank Tomomichi Amano, Eric Anderson, Yuxin Chen, Pradeep Chintagunta, Jörg Claussen, J. P. Dubé, Jun Kim, Pedro Gardete, Brett Gordon, Wes Hartmann, Günter Hitsch, Ayelet Israeli, Anja Lambrecht, Mitch Lovett, Carl Mela, Blake McShane, Harikesh Nair, Sridhar Narayanan, David Reiley, Navdeep Sahni, Brad Shapiro, Ken Wilbur, Florian Zettelmeyer, Juanjuan Zhang, and seminar participants at the London Business School, INSEAD, Duke University, Temple University, the University of British Columbia, the University of California at San Diego, the University of Colorado, the University of Florida, the University of Minnesota, and Stanford Graduate School of Business, as well as conference participants at the 2016 Workshop on Economics of Advertising and Marketing, the 2016 Greater-China Conference on Mobile Big Data Marketing, the 2015 Big Data Conference, Searle Conference on Internet Commerce and Innovation, the 2015 Marketing Science Conference, the 2015 Quantitative Marketing and Economics Conference, and the Yale China India Insights Conference for valuable comments. The authors thank Ipetram Inc., Zhanbo Zhao, and Haoxin Lu for their help with data collection. The authors also thank Audrey Siyu Zhao for being an excellent research assistant. All errors are the authors.

Endnotes

[1] One notable exception is Lovett and Staelin (2016), who estimate a structural model that allows for complementarity between WOM and TV viewing.

[2] The meta-analysis of You et al. (2015) finds an average elasticity of 0.2. For the subset of elasticities that share characteristics with our study (e.g., private consumption), the average elasticity is roughly equal to 0.6. Lovett and Staelin (2016) estimate a more moderate elasticity of 0.04.

[3] Our comparison is based on the effectiveness of "per unit of WOM" and does not take the cost of providing WOM into account. Despite a lower elasticity, WOM could compare favorably in terms of return on investment to TV advertising, because of cost differences.

[4] We define the number of comments pertaining to a specific episode of a show as the total number of relevant comments posted on the day the episode aired.

[5] Specifically, we control for show/weekday pair fixed effects, holidays, and the duration of the world team table tennis championship (sports events are not included in the set of shows in our estimation sample).

[6] We note that our panel of shows is unbalanced because most shows do not air every day. The time series of average ratings is therefore affected by compositional changes in shows over time.

[7] In addition, a significant literature studies various aspects of interactions of users within social media platforms, but does not analyze the impact on consumers' purchase decisions. Trusov et al. (2010) propose a model to identify influential users on social media websites. Zhang and Zhu (2011), Aaltonen and Seiler (2016), Shriver et al.

(2013), Toubia and Stephen (2013), and Ahn et al. (2016) investigate why users contribute content to social network platforms such as Wikipedia and Twitter. Ameri et al. (2016) estimate the impact of observational learning and WOM on users' adoption decisions on an anime network.

[8] We exclude local channels from this calculation.

[9] The block was in place from 8 A.M. on March 31 to 8 A.M. on April 3. Because we analyze the effect of microblogging on prime-time TV shows that air after 6 P.M., and because most relevant tweets and comments occur after 8 A.M., we treat the three calendar days March 31 to April 2 as the time period of the block.

[10] The market share for each episode is defined as the share of households watching the particular episode at the time it airs among all households that own a TV. Because many consumers do not watch any TV at a given point in time, the rating numbers are generally quite low. The ratings only measure live TV viewership and do not include delayed viewing.

[11] Retrieved from Weibo's SEC filings (http://ir.weibo.com/phoenix.zhtml?c=253076&p=irol-sec, accessed November 2015) in April 2014. The daily numbers of users and tweets are calculated based on data from December 2013.

[12] One can also comment on a retweet, but we consider only the comments on the "original" tweet.

[13] During the time period of our data collection, Sina Weibo altered the algorithm that displays historical tweets, and currently displays only a subset of the full set of relevant tweets for a specific keyword. Most of our data collection was already completed when this change happened. However, we collected some pieces of data after the change, and we therefore need to rescale this data to reflect the fact that Sina Weibo now displays only a subset of all tweets. Section A of the online appendix provides more details on how we handled this issue.

[14] We use the time stamp of each tweet to define whether the tweet falls into the relevant time window. For all other types of activity, we use the time stamp associated with the tweet to which they belong. Other types of activity typically occur within a short time window after the original tweet. Specifically, we find that more than 50% of the comments were posted within 41 minutes of the original tweet. (To establish this pattern, we analyzed a sample of 12,000 comments for which we collected time stamps.)

[15] The government gave no official statement linking the Weibo block explicitly to political events, but news sources clearly connected them (see, e.g., Johnson 2012).

[16] Another Chinese microblogging website, Tencent Weibo, also blocked the comment function during the same period. We focus on Sina Weibo because Tencent Weibo had only about 9% of the market share, whereas Sina Weibo held nearly 90%. The authorities also closed down 16 minor websites for spreading political rumors.

[17] This Chinese word translates to "jumping over the wall."

[18] We also attempted to collect data on the set of other sources that indicate an interest in the scandal, which we discussed in Section 3 (see, in particular, the last paragraph in that section) by geography. However, for most variables, the data are quite sparse when cutting by geography. For example, "Gu Kailai" and "U.S. Consulate" had a volumes of Baidu searches and barely appeared in Google news articles in Hong Kong. We therefore only include graphs for the two additional keywords for which we are able to obtain sufficient data, "Bo Guagua" and "Chongqing." The correlations between Hong Kong and mainland China for those two search queries are equal to 0.86 and 0.82, respectively. The graphs of these time series are presented in Figure A6 in the online appendix. Generally, we would like to compare other sources as well, such as online discussions of the scandal in the two geographies. However, because such discussions were censored in mainland China, we cannot implement such a comparison.

[19] We report the rating distribution for mainland shows in both Shenzhen and Hong Kong in the lower panel of Table A3 in the online appendix.

[20] In a similar spirit, we could, in principle, also analyze shows of Hong Kong origin that air in Hong Kong as well as in Shenzhen. However, only two of the six Hong Kong channels can be received in Shenzhen, and the market share of those shows is small in Shenzhen. Furthermore, Hong Kong shows are primarily targeted at local consumers, and hence they are not talked about much on Sina Weibo. Therefore, the small subset of consumers in Shenzhen that watch the shows are unlikely to use Sina Weibo to inform themselves and talk about these shows.

[21] We define Weibo usage as the number of times the three keywords were searched in a specific city divided by the population size of the respective city. Because the scale of this variable is hard to interpret, we normalize the variable to lie between zero and one.

[22] We ran the same regression using Internet penetration rates as a proxy for Sina Weibo usage at the local level and found similar results.

[23] Results are similar when grouping shows according to only their preblock level of activity. However, a small set of shows (14) are not observed before the block (they started airing during the block). Therefore, using postblock data allows us to assign an activity level to that set of shows. We report descriptive statistics for the average number of comments per episode (outside the block) in the lower panel of Table 1.

[24] Consumers could conceivably substitute away from high- toward lower-activity shows. This type of substitution between shows does not seem to occur, because none of the groups of shows captured in the regression experiences an increase in ratings. In Section E in the online appendix, we further explore whether consumers substitute between shows or to the outside option of not watching any TV. We find that comments primarily lead consumers to substitute toward the outside option rather than other shows.

[25] We include interactions with an exhaustive set of genre dummies in this regression and thus do not include the censor dummy without interaction.

[26] We also note that although we find a significant effect only for news shows, our results are robust to dropping news shows from the sample entirely. When estimating Equation (1) without news shows, we obtain a coefficient (standard error) on the censor dummy of −0.016 (0.007). By comparison, the coefficient (standard error) based on all shows is equal to −0.017 (0.005).

[27] In each case, the show characteristic is defined as a binary distinction. We hence run our baseline regression with the censor dummy as well as an interaction of the dummy with a dummy for one of the binary values in each case. The coefficients (standard errors) in these cases are as follows: interaction of the censor dummy with the new-show dummy, 0.001 (0.013); interaction with the daily-frequency dummy, −0.010 (0.013).

[28] The coefficient (standard error) on the interaction of the rerun dummy with the block dummy is equal to 0.029 (0.010).

[29] Three DiD regressions across geographies each contain one relevant interaction term, whereas the across-show regression contains two relevant interactions (because we are considering three different levels of Weibo activity), giving us a total of five relevant coefficients across all four regressions. For each regression, we implement two placebo tests and hence obtain a total of 10 relevant coefficients.

[30] To mirror our regression specification, we first regress ratings on show and weekday fixed effects and then compute the average rating residual for each show and time period. We also weigh shows by the number of episodes aired during the relevant time period when computing average ratings.

[31] We do not apply this approach to the regression based on an interaction with the (continuous) Weibo penetration rate variable with the

censor dummy, because no binary distinction exists between treatment and control groups in that specification.

[32] The *p*-values for the three main specifications (mainland China/Hong Kong, low-/high-activity shows, cities with low/high Sina Weibo penetration) are equal to 0.012, 0.012, and 0.077 when collapsing the data, relative to 0.028, 0.026, and 0.006 when clustering standard errors at the show level.

[33] We also include a dummy for whether an episode was actually aired on the specific day.

[34] In the top panel of Table 4, we report the standard deviation of the residuals when regressing the dependent variable in the respective column onto show fixed effects. This metric reflects the amount of variability over time after controlling for cross-sectional differences across shows.

[35] Because of the (weak) uptake of retweets during the block, the effect of ratings per unit of comment might be larger than our estimate.

[36] This result is closely related to the regressions of log ratings on the block dummy presented in Section 4. In particular, column (1) of Table 2 represents the reduced form of the IV regression presented above. Hence, the multiplication of the first-stage coefficient of the block dummy and the second-stage coefficient on log comments is equal to the coefficient on the block dummy in column (1) of Table 2.

[37] An alternative way to assess the economic relevance of the estimated effect is to consider the impact of microblogging on revenue from TV advertising (which is tied to ratings). We implement such a calculation in Section C of the online appendix and find that a one-standard-deviation shift in the number of comments (76,000 comments) entails a change in advertising revenue per episode of US$190,000.

[38] See *Wikipedia*, s.v. “1% rule (Internet culture),” retrieved June 6, 2016, https://en.wikipedia.org/wiki/1%25_rule_(Internet_culture).

[39] We provide more detail on the sampling process in Section D in the online appendix.

[40] In Section 4.2, we used the total number of comments throughout the day on which a specific show aired. We now categorize comments with regard to whether they were posted before, during, or after the episode aired. We primarily focus on pre- and postshow comments, but also investigate the impact of comments posted during the show. When including during-show comments in addition to pre- and postshow comments, we find during-show comments have no significant impact on ratings, and the coefficients on pre- and postshow comments are similar to those in the specification without during-show comments presented in Table 5.

[41] Similarly, the difference in characteristics between shows with low, medium, and high levels of preshow comments is also small, albeit slightly more pronounced. The same is true when comparing characteristics across shows with low, medium, and high levels of postshow comments.

[42] When including lagged terms, the regressions contain multiple endogenous regressors; therefore, a separate first stage (and a separate *F*-statistic) exists for each of the endogenous variables. We report the *F*-statistics for all excluded instruments for each of the first-stage regressions in each case.

References

Aaltonen A, Seiler S (2016) Cumulative growth in user generated content production: Evidence from Wikipedia. *Management Sci.* 62(7):2054–2069.

Ahn D-Y, Duan JA, Mela CF (2016) Managing user-generated content: A dynamic rational expectations equilibrium approach. *Marketing Sci.* 35(2):284–303.

Ameri M, Honka E, Xie Y (2016) Word-of-mouth, observational learning, and product adoption: Evidence from an anime platform. Working paper, University of Texas at Dallas, Dallas.

Anderson M, Magruder J (2012) Learning from the crowd: Regression discontinuity estimates of the effects of an online review database. *Econom. J.* 122(563):957–989.

Aral S, Walker D (2014) Tie strength, embeddedness, and social influence: A large-scale networked experiment. *Management Sci.* 60(6):1352–1370.

Bagwell K (2007) The economic analysis of advertising. Armstrong M, Porter R, eds. *Handbook of Industrial Organization*, Vol. 3 (North-Holland, Amsterdam), 1701–1844.

Becker GS, Murphy KM (1993) A simple theory of advertising as a good or bad. *Quart. J. Econom.* 108(4):941–964.

Bertrand M, Duflo E, Mullainathan S (2004) How much should we trust differences-in-differences estimates? *Quart. J. Econom.* 119(1):249–275.

Blake T, Nosko C, Tadelis S (2015) Consumer heterogeneity and paid search effectiveness: A large scale field experiment. *Econometrica* 83(1):155–174.

Chevalier JA, Mayzlin D (2006) The effect of word of mouth on sales: Online book reviews. *J. Marketing Res.* 43(3):345–354.

Chintagunta PK, Gopinath S, Venkataraman S (2010) The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing Sci.* 29(5):944–957.

Dhar V, Chang EA (2015) Does chatter matter? The impact of user-generated content on music sales. *J. Interactive Marketing* 23(4):300–307.

Gilchrist DS, Sands EG (2016) Something to talk about: Social spillovers in movie consumption. *J. Political Econom.* 24(105): 1339–1382.

Godes D, Mayzlin D (2004) Using online conversations to study word-of-mouth communication. *Marketing Sci.* 23(4):545–560.

Gong S, Zhang J, Zhao P, Jiang X (2017) Tweeting as a marketing tool—Field experiment in the TV industry. *J. Marketing Res.* Forthcoming.

Gordon B, Zettelmeyer F, Bhargava N, Chapsky D (2016) A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. Working paper, Northwestern University, Evanston, IL.

Gordon BR, Hartmann WR (2013) Advertising effects in presidential elections. *Marketing Sci.* 32(1):19–35.

Hartmann WR, Klapper D (2017) Super Bowl ads. *Marketing Sci.*, ePub ahead of print October 5, https://doi.org/10.1287/mksc.2017.1055.

Johnson I (2012) Coup rumors spur China to hem in social networking sites. *New York Times* (March 31). https://nyti.ms/2pTZsAz.

Lambrecht A, Tucker C, Wiertz C (2017) Should you target early trend propagators? Evidence from Twitter. *Marketing Sci.* Forthcoming.

Lewis RA, Rao JM (2015) The unfavorable economics of measuring the returns to advertising. *Quart. J. Econom.* 130(4):1941–1973.

Lewis RA, Reiley DH (2014) Online ads and offline sales: Measuring the effect of retail advertising via a controlled experiment on Yahoo! *Quant. Marketing Econom.* 12(3):235–266.

Liu Y (2015) Word of mouth for movies: Its dynamics and impact on box office revenue. *J. Marketing* 70(3):74–89.

Lovett MJ, Staelin R (2016) The role of paid, earned, and owned media in building entertainment brands: Reminding, informing, and enhancing enjoyment. *Marketing Sci.* 35(1):142–157.

Luca M (2016) Reviews, reputation, and revenue: The case of Yelp.com. Working paper, Harvard Business School, Cambridge, MA.

Petrova M, Sen A, Yildirim P (2016) Social media and political donations: Evidence from Twitter. Working paper, Universitat Pompeu Fabra, Barcelona, Spain.

PR Newswire (2013) Marketers say “word of mouth marketing” is more effective than traditional marketing; forecast big increase in social media spending. (November 19), http://www.prnewswire.com/news-releases/marketers-say-word-of-mouth-marketing-is-more-effective-than-traditional-marketing-forecast-big-increase-in-social-media-spending-232486271.html.

Rossi PE (2014) Invited paper—Even the rich can make themselves poor: A critical examination of IV methods in marketing applications. *Marketing Sci.* 33(5):655–672.

Sahni N (2015) Effect of temporal spacing between advertising exposures: Evidence from an online field experiment. *Quant. Marketing Econom.* 13(3):203–247.

Sethuraman R, Tellis GJ, Briesch RA (2011) How well does advertising work? Generalizations from a meta-analysis of brand advertising elasticities. *J. Marketing Res.* 48(3):457–471.

Shapiro BT (2016) Positive spillovers and free riding in advertising of prescription pharmaceuticals: The case of antidepressants. *J. Political Econom.* Forthcoming.

Shriver SK, Nair HS, Hofstetter R (2013) Social ties and user-generated content: Evidence from an online social network. *Management Sci.* 59(6):1425–1443.

Sinkinson M, Starc A (2015) Ask your doctor? Direct-to-consumer advertising of pharmaceuticals. Working paper, University of Pennsylvania, Philadelphia.

Sonnier GP, McAlister L, Rutz OJ (2011) A dynamic model of the effect of online communications on firm sales. *Marketing Sci.* 30(4):702–716.

Toubia O, Stephen AT (2013) Intrinsic vs. image-related utility in social media: Why do people contribute content to Twitter? *Marketing Sci.* 32(3):368–392.

Trusov M, Bodapati A, Bucklin RE (2010) Determining influential users in Internet social networks. *J. Marketing Res.* 47(4):643–658.

Trusov M, Bucklin RE, Pauwels K (2009) Effects of word-of-mouth versus traditional marketing: Findings from an Internet social networking site. *J. Marketing* 73(5):90–102.

Tuchman A (2016) Advertising and demand for addictive goods: The effects of e-cigarette advertising. Working paper, Northwestern University, Evanston, IL.

Tuchman A, Nair HS, Gardete P (2017) An empirical analysis of complementarities between the consumption of goods and advertisements. *Quant. Marketing Econom.* Forthcoming.

Tucker CE (2014) Social networks, personalized advertising, and privacy controls. *J. Marketing Res.* 51(5):546–562.

Villanueva J, Yoo S, Hanssens DM (2008) The impact of marketing-induced versus word-of-mouth customer acquisition on customer equity growth. *J. Marketing Res.* 45(1):48–59.

You Y, Vadakkepatt GG, Joshi AM (2015) A meta-analysis of electronic word-of-mouth elasticity. *J. Marketing* 79(2):19–39.

Zhang XM, Zhu F (2011) Group size and incentives to contribute: A natural experiment at Chinese Wikipedia. *Amer. Econom. Rev.* 101(4):1601–1615.

Zhu F, Zhang XM (2010) Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *J. Marketing* 74(2):133–148.