For Whom to Tweet? Evidence from a Large-Scale Social Media Platform[*]

Zaiyan Wei
Krannert School of Management, Purdue University, West Lafayette, IN 47907, zaiyan@purdue.edu

Mo Xiao
Eller College of Management, University of Arizona, Tucson, AZ 85721, mxiao@eller.arizona.edu

January 2017

We study the effects of peer-group sizes on tweeting in a large-scale and influential social media platform. Tweets in social media disseminate information and exert social influence. However, 50% of the users post less than 6 tweets per month and contribute to less than 15% of the tweets in stock, while the top 10% post over 40 tweets a month and contribute to more than half of the tweets in stock. We attribute the highly unbalanced contribution to a user's conflicting incentives of free-riding and maximizing social influence. We exploit the asymmetry of a user's peer groups (followers and followees, groups of people following and being followed by the user) to disentangle these incentives, and devise empirical strategies to deal with the endogenous network formation. We find asymmetric effects, in both signs and sizes, of followers and followees. A larger group of followers leads a user to tweet more, while a larger group of followees leads a user to tweet less. As the follower effects are dominant, our simulations indicate that by randomly adding 1% new connections the platform could increase the total tweets by 25%. Targeting occasional tweeters is even more effective in promoting the activeness of this platform.

Key words: social media and social networks, user-generated content, peer effects, public goods

[*] We would like to thank Keisuke Hirano, Mingfeng Lin, Yong Liu, Mohammad Rahman, Stanley Reynolds, Jin Wang, as well as participants of 2016 Annual ASSA/AEA Conference, 2016 International IO Conference, and 2016 INFORMS Conference on Information Systems and Technology for comments and suggestions.

1. Introduction

Internet-enabled social media platforms permeate our social and economic lives. These platforms largely depend on the contributions of individual users. The user-generated content has unambiguously significant impacts on many aspects of business such as branding and advertising (Gopinath, Chintagunta, and Venkataraman 2013, Kumar et al. 2013, Sun and Zhu 2013), consumer learning and response (Zhao et al. 2013, Gans, Goldfarb, and Lederman 2016), product demand and sales (Zhu and Zhang 2010, Ma, Krishnan, and Montgomery 2015, Gong et al. 2016, Kumar et al. 2016), employee productivity (Wu 2013, Huang, Singh, and Ghose 2015), and even the market value of firms (Nam and Kannan 2014, Schweidel and Moe 2014).[1] Services provided by these social media platforms are widely adopted by users across age, gender, income, and ethnicity. As the largest social media platform, Facebook.com had attracted more than 1.65 billion registered users by April 2016. Twitter.com had surpassed 310 million active users per month by the first quarter of 2016. In China, 47% of its population, or more than 600 million people, use social media platforms.[2]

Social media platforms are often equipped with social networking services (SNS). When generating or sharing content on these platforms, users are influenced by their “neighbors” (Zhang, Liu, and Chen 2015). Particularly, the size of peer groups has a prominent impact on the content generating process (Zhang and Zhu 2011). Different from the structure of bilateral friendship networks that are prevalent on platforms such as Facebook.com, in Twitter-type networks, a user’s peer groups include followees (users she follows) and

[1] The influence of social media goes beyond business. It has broad impact on nearly every aspect of our social lives and the society. For example, social media usage is shaping the workings of democracy globally (Chi and Yang 2011). In particular, Twitter.com is believed to be the “heartbeat” of 2016 US Presidential election (Manjoo 2016).

[2] PewResearch Internet Project annual reports from online surveys can be found at `http://www.pewinternet.org/fact-sheets/social-networking-fact-sheet/`. Annual survey by “We Are Social” reports the statistics of online social networks: `http://wearesocial.net/blog/2015/01/digital-social-mobile-worldwide-2015/`.

followers (users following her). On one hand, a user can read all her followees' tweets, and so the followees constitute her major source of information in the network; on the other hand, her tweets can be viewed by all her followers. Every user is keenly aware of this fact. Previous studies have examined the effects of a user's followers (Toubia and Stephen 2013) or the effects of overlap in online neighbors (Peng et al. 2016); we seek to fill the gap in the literature by separating the effects of followees and those of followers on content generation and sharing.

Separating the followee effects and the follower effects may be key for us to understand the highly skewed distribution of user-generated content in such a social platform. In the large-scale network we observe, approximately 10% of the users contribute more than 50% of the total tweets in stock, while half of them provide less than 15% of the tweets. Similarly, 50% of the users post fewer than six tweets monthly, while a user at the 90th percentile generates on average 40 tweets a month. For a platform seeking to promote its prosperity, and a commercial user seeking to manage its brand, and solicit feedback and sales from (potential) consumers, the lack of participation from the majority of the users on the platform presents a problem. *In this paper, we endeavor to understand and remedy this problem by looking into the tweeting incentives of a noncommercial user, in particular, "for whom to tweet?"*

We exploit the asymmetry of Twitter-type networks to disentangle a user's conflicting incentives of free-riding and social benefit considerations when providing tweets as a public good. We view user-generated content as public good provision on the Internet (Duan, Gu, and Whinston 2008). Undersupply is what typically characterizes public good provision due to free-riding incentives (Olson 1965). In a "directed" network, the information embedded in tweets passes on mainly from a user to her followers, although the reverse direction of information transmission is also possible.[3] A user may free ride on others' (including followee and follower) contribution. With an increase in one's peer group, a user may think a relevant piece of information is already disseminated among the shared base of followers so there is less need for her to exert effort (to provide the public good). That is, the free-riding incentive postulates the negative effects of both the followee count and the follower count.[4] Social benefit considerations, however, may counteract the free-riding incentives. Social benefits could be purely altruistic motives (Arrow 1972, Chamberlin 1974), seeking social influence (Becker 1974), or motivated by the perception of others (Fehr and Falk 2002). The common theme of these benefits is that they increase with a larger audience; as a result, the size of both followees and followers may have *positive* effects on one's tweeting.[5] Summing up, the overall effects of the followee count or the follower count on one's tweeting are not *a priori* clear cut. A negative effect of the followee/follower count indicates that the free riding incentives dominate the social benefit incentives; a positive effect indicates otherwise.

We implement this idea using data from Tencent Weibo, Twitter's counterpart in China. Tencent Weibo is the largest Twitter-type social media platform in China, with over 220 million registered users by 2014. The main challenge in identifying the effects of asymmetric group size is the endogeneity of network formations (Manski 1993). In our context, the unilateral relationships of following are established by users who deliberately choose to connect. Econometrically, there exist user unobserved characteristics correlated with

[3] Although one's tweets appear automatically on all her followers' home pages, her followers' tweets do not necessarily show up on her page. Unless a follower is simultaneously her followee or she clicks on the follower's tweets page intentionally, she does not receive updates automatically.

[4] In addition, the negative effect of followers count may reflect a user's privacy concern. With more users following her, particularly random accounts in the network, she would be more cautious about tweeting to the public. Therefore she is more selective of what to tweet, and would not tweet as much as she used to.

[5] A user's followees count may have positive effect on her tweeting merely because the amount of information acquired is positively correlated with the number of followees.

network formations that in turn determine the network sizes, and these characteristics simultaneously determine her tweeting behavior.

We address the endogeneity problem by devising different identification strategies using two datasets from Tencent Weibo. The first one we obtain is a proprietary longitudinal dataset containing over $100,000$ observations of $20,289$ Weibo users over a four-month period between August 2011 and December 2011. To alleviate the endogeneity problem, we consider a panel approach that incorporates user and time fixed effects to control for unobserved user heterogeneity that is constant over time, as well as other unobservables that affect all users but differ in time.

To further deal with the problem that there may exist time-varying unobservables, we obtain a second dataset, which is a random sample of the snapshot of the whole network with a much larger size and a more complete network structure. This administrative sample contains a cross section of roughly 1.4 million users up to a date in January 2012, with their activities and all followee and follower identities. We utilize the network structure to propose an instrumental variable (IV) method. Specifically, we use the average observed characteristics (including age and gender) of one's second-order followees and those of second-order followers as instruments for all endogenous variables. Under the assumption that a Weibo user does not tweet to win over her second-order neighbors (Murthy 2012, Lee, Hosanagar, and Tan 2015),[6] the characteristics of the second-order neighbors affect her first-order neighbors' tweeting activities, which in turn affect her utilities of forming connections. Therefore, the IVs are correlated with the user's number of first-order neighbors, but independent of the characteristics governing her tweeting.

[6] Murthy (2012) argues that the audience range of posts on Twitter-type social media is typically larger than the perceived range, although tweet posters often intend to circulate to their immediate followers instead of higher-order audience. Lee, Hosanagar, and Tan (2015) study an interesting application in the context of online product ratings and find that friends' ratings have significantly larger effects than the ratings by the crowd.

We find that, from both the longitudinal and cross-sectional analysis, a larger number of followees leads to fewer tweets while a larger number of followers has the opposite effects. We further establish the relative magnitude of these two effects. We find that the (positive) follower effect is much stronger than the (negative) followee effect. That is, the positive effect of followers may be (partly) offset by the negative effect of followees. As such, user recommendations may lead to ambiguous results because of this trade off. To analyze this trade off, we perform simulations that randomly generate 1% new connections. As the (positive) follower effect strongly dominates the (negative) followee effect, we find that this 1% new connections will increase the total tweets by 25% on this particular platform. Targeting less "active" users (in another set of simulations we randomly generate followers to targeted users—those with *fewer* followers) will increase the volume of tweets even more.

Similar to the findings for the volume of tweets, we also find asymmetric effects on a user's binary decision whether to tweet at all during a certain period. From our longitudinal data, we find that a user is significantly less likely to engage in (any) tweeting with a larger group of followees, while the follower effect is not significant. These findings complement the literature that in addition to the positive effects of follower count, there exist negative impacts of a user's followees. Going beyond these average effects, are the impacts different for users with different characteristics? To answer this question, we examine the moderating roles of user characteristics in the asymmetric effects summarized earlier. Interestingly enough, we find that the follower effect is more salient for users with more "active" followers, who generate more tweets on average, and that neighbors sizes (both follower and followee) have a greater impact on female users.

Our work uses the data that include the structure of the user's social networks to distinguish the asymmetric influences of followees and followers in directed networks. We add to the strand of the literature studying network effects in social media (Aral and Walker 2011, Bond et al. 2012, Wu 2013, Ma, Krishnan, and Montgomery 2015), in particular, in directed networks. The closest to our study is Toubia and Stephen (2013). They conducted a field experiment that added followers randomly to a treated group of Twitter users to study the effects of follower count on a user's incentives to tweet. We take a step further by showing that not only the number of followers but also followees affect tweeting decisions significantly and that these effects are opposite to each other in both signs and magnitude. Our approach enriches our understanding of a directional network, suggesting that we need to consider the integrated whole of a network instead of only its "parts." In another work, Peng et al. (2016) study how the common followees and common followers (between tweet senders and receivers) affect users' content sharing behavior—retweeting. In contrast, we complement this line of research by systematically investigating how the two peer groups (asymmetrically) influence both content sharing and generation in social media.

We also contribute to the broader literature on the provision of public goods in social networks (Olson 1965, Andreoni 2007, Bramoullé and Kranton 2007, Chen et al. 2010). Different from traditional social networks, we study a virtual network where the networks of people may not have any real world connections. We show that these virtual networks have real impact on individuals' decisions to speak on line. Our results have managerial implications for both platforms promoting the activeness of their users and commercial users running marketing efforts on these platforms. We recommend strategies that can create a more leveled playground. Our findings of asymmetric effects of followees and followers suggest marketers need to exert greater caution in attracting on-site followers and following noncommercial users.

2. Tencent Weibo

Tencent Weibo is the largest microblogging platform in China and provides Twitter-type social networking services. It was launched on April 1, 2010 as an affiliated website to Tencent.com.[7] By January 2014, over 220 million users had registered on the platform, and the average number of daily active users had reached a record of 100 million.

Users of Tencent Weibo can either "tweet" or "retweet" other users' tweets. A *tweet* is a short note comprised of texts, links, or graphics that a user posts on her public profile page; while a *retweet* is a re-post or sharing of other users' tweets, also on her profile page. We characterize tweets (and retweets) as public goods, since they are free to access within the community. The maximum length of a tweet (as well as a retweet) is 140 Chinese characters. In addition to tweeting, Weibo users communicate with each other by *tagging* others in a tweet, *commenting* below a tweet,[8] or sending *messages* privately.

In addition to providing the micro-blogging service summarized above, Weibo users form a *directed* network that is connected by the "following" relationship. A Weibo user can choose to follow another user's tweets without the other party's consent. In this relationship, we define the focal user the "follower" and the other party as the "followee." A connection can be established unilaterally (meaning that the followee does not have to follow the follower's tweets); therefore, the Weibo network is directed. Once a following relationship is established, all followees' tweets and retweets automatically show up on the follower's page. It is, however, asymmetric that the follower's tweets do not appear on the followee's pages. With this particular feature, the followers are also called a user's "audience" on Tencent Weibo. We provide comparisons with Twitter.com in Appendix A.

[7] Tencent is the largest networking service provider in China. The services they provide include instant messaging, personal space, and micro-blogging. Tencent's flagship products are QQ and WeChat. By the end of 2013, there had been over 0.8 billion registered users on Tencent; they had reached a peak of 0.18 billion users online simultaneously. Their official website provides more information: `http://www.tencent.com/en-us/index.shtml`.

[8] More specifically, a user can mention others, not necessarily her followees or followers, in certain tweets (or comments) by typing their user names following the "@" symbol directly in the texts. Any user can post comments below a tweet unless the setting is customized. The comments, together with the tweets, are listed publicly on Weibo pages. Any registered user can read the content.

3. Data and Samples

We obtained two datasets from Tencent.com. The first one is a proprietary dataset consisting of $29,956$ Weibo users with their Tencent identities,[9] tweeting, and networking information between August 2011 and December 2011. Specifically, Tencent kept track of these users' number of followees and followers,[10] the number of tweets, retweets, and other activities such as comments and private messages. The data also include users' demographic information including location and job information verification status.

The specific sampling process of the longitudinal dataset is as follows. It contains a sample of active users who registered before August 1, 2011. Tencent reported their records at six different dates between August 15, 2011 and December 11, 2011. From this dataset, we construct a sample of $20,289$ users with complete records at each of the six observation dates, *i.e.*, constituting a balanced panel. This is the main sample used in our estimations.[11] Table 1 summarizes this longitudinal sample. Particularly, 50% of the users post fewer than 6 tweets a month, while a user at the 90th percentile generates about 40 tweets a month. That is, even a median user is a member of the "silent majority."

We obtained another dataset from Tencent.com. It contains a snapshot of a much larger set of users, including data on network structure. Tencent.com publishes a random sample of $1,392,873$ users from its Weibo user pool on a website.[12] For these users, Tencent provides their networks, tweeting, and demographic information up to a date in January

[9] Each user has a unique identity on Tencent.com. The identity is a number with 5 to 11 digits. Users can use any Tencent service with this identity, called "QQ Number."

[10] Unfortunately, Tencent does not provide the identifies of these followees or followers so we cannot construct (at least part of) the network.

[11] We focus on the balanced panel mainly because the data provider failed to collect the missing records, but not because those records do not exist or are randomly missing. We conduct a robustness check using the original sample of $29,985$ users. Results are consistent with those from our main sample.

[12] Tencent.com hosted the 2012 Knowledge Discovery and Data Mining (KDD) Cup competition in 2012. This is an annual global competition of data mining that targets all data scientists all over the globe. The competition website is `https://www.kddcup2012.org/`. In one of the two tracks, Tencent made friendship connecting recommendations to the $1,392,873$ Weibo users and kept track of their decisions over a period of time. The task for participants was to develop statistical models to predict user decisions of accepting or rejecting the recommendations.

Table 1 Summary Statistics of the Panel Data

	Total Tweets[a]			**New Total Tweets**		
Dates	Med.	Mean	s.d.	Med.	Mean	s.d.
2011-08-15	25	74.014	188.054	-	-	-
2011-09-15[b]	36	96.222	221.432	7	22.208	56.191
2011-10-31	40	104.361	247.062	2	8.139	48.340
2011-11-15	42	110.165	268.005	1	5.804	33.710
2011-11-30	47	119.694	299.714	2	9.529	48.046
2011-12-11	49	123.502	318.153	1	3.808	27.998
	# Followees			**# Followers**		
Dates	Med.	Mean	s.d.	Med.	Mean	s.d.
2011-08-15	20	34.428	746.238	34	43.960	68.083
2011-09-15	25	39.269	724.470	36	46.994	70.986
2011-10-31	29	47.167	774.478	39	50.440	76.163
2011-11-15	31	49.884	788.067	40	51.675	78.252
2011-11-30	33	52.511	818.095	41	52.697	77.944
2011-12-11	34	55.223	876.927	41	53.888	80.157
Users			20,289			

[a] The total number of tweets is the sum of tweets and retweets, not including other activities such as comments and messages.

[b] For this cross-section, the network sizes were collected on September 15, 2011, while the tweeting variables were recorded on October 10, 2011. We conduct a robustness check by excluding this cross-section. Results are qualitatively the same as our main empirical findings.

2012 (the exact date not revealed by Tencent). Particularly, the dataset contains the *coded* identities (different from their "QQ numbers"—users' unique identities on Tencent) of these users' followees and followers. For all original 1,392,873 users and their immediate followees and followers, the dataset contains all their number of tweets, retweets, other activities such as comments, and demographic information including age and gender. Compared with the panel data, this cross-sectional dataset documents a larger sample and more complete network structure with identities of all followees and followers, although it does not keep track of higher-order neighbors. As we observe the identity number of all neighbors, we are able to construct at least parts of the high-order neighbors.

From the original random sample, we construct a set of 402,470 users[13] as the main sample in our estimations. These users have at least one second-order followee and one

[13] In Appendix C we compare this constructed sample with the original dataset consisting of 1,392,873 users. We find that this sample is representative of the original dataset in terms of demographic information.

second-order follower. As mentioned in the introduction, the instruments we propose are the average observed characteristics of second-order neighbors. This is the reason we focus on this subset of users. Table 2 and 3 summarize this cross-sectional sample. We notice that 50% of the users have less than 19 followees and 2 followers, while a user at the 90th percentile has 93 followees and 5 followers. We also observe a highly skewed distribution of tweets from the sample. Although we worry that the observations are quite noisy because the standard deviations are in the order of magnitude larger than either the mean or median, we think the large standard deviation is due to the highly skewed distribution of tweets, which is the nature of this type of social network. Part of the goal of this paper is to devise mechanisms to encourage more active participation of low percentile tweets and generate a more even distribution of tweet behavior. Later, we also carry out robustness checks by excluding "VIP" users with a huge number of followers or followees in our empirical analysis later.

Comparing Table 1 with Table 2 we notice that users in the panel data have fewer tweets (the median of the tweets distribution is 49 by the end of the sampling period) than those sampled in the cross-sectional data (with the median 143 by January 2012). In contrast, half of the users in the cross-sectional sample have more than 19 followees and 2 followers, while half of the users had at least 34 followees and 41 followers before January 2012, which is the sampling date of the cross-sectional data. As the cross-sectional data represent a random sample from the Weibo user pool, the panel data may have oversampled users with larger number of neighbors.

4. Empirical Strategies and Results

A major challenge of identification is that a user's network or peer group—followees and followers on Weibo—is endogenously formed (Manski 1993, Bramoullé, Djebbari, and Fortin

Table 2 Summary Statistics of the Cross-Sectional Data

	Summary Statistics				
Variables	Med.	Mean	s.d.	Min.	Max.
# Total Tweets[a]	143	256.913	447.180	0	65,518
# Tweets	120	208.704	371.798	0	65,506
# Retweets	4	48.209	177.289	0	21,780
# Comments	1	6.458	36.653	0	13,384
Network Information					
# Followees	19	42.427	80.475	1	5,188
# Followers	2	17.203	692.496	1	159,453
# Second-Order Followees	623	935.278	1,162.658	1	38,424
# Second-Order Followers	4	845.956	13,683.390	1	1,195,098
# All Neighbors	22	59.631	698.584	2	159,496
# Friends[b]	1	1.522	5.671	0	454
Log-Transformations					
log (**# Total Tweets** + 1)	4.970	4.842	1.288	0	11.090
log (**# Followees**)	2.944	2.918	1.292	0	8.554
log (**# Followers**)	0.693	0.679	0.913	0	11.980
Observed Characteristics					
Age	22	24.123	17.147	0	123
1(Missing Year Birth)	0	0.014	0.117	0	1
1(Female)	1	0.511	0.500	0	1
1(Missing Gender)	0	0.007	0.085	0	1
1(Age Outliers)[c]	0	0.094	0.291	0	1
Users			402,470		

[a] The total number of tweets is the sum of tweets and retweets.

[b] A "friend" is a user who is both an individual's followee and follower.

[c] Defines users with year of birth being before 1940 or after 2000 as "outliers."

Table 3 Summary Statistics of the Cross-Sectional Data—Peer Groups

	All Neighbors		**2nd-Order Followees**		**2nd-Order Followers**	
Variables[a]	Mean	s.d.	Mean	s.d.	Mean	s.d.
# Total Tweets	540.086	443.226	595.012	213.273	431.540	516.571
Age	30.186	7.051	32.401	3.066	25.302	11.251
1(Missing Year Birth)	0.071	0.075	0.076	0.039	0.011	0.071
1(Female)	0.409	0.188	0.351	0.089	0.451	0.331
1(Missing Gender)	0.063	0.067	0.077	0.031	0.006	0.051
1(Age Outliers)	0.149	0.111	0.167	0.045	0.093	0.180
Users				402, 470		

[a] For each variable, we calculate the average value among the corresponding peer group for each user. This table reports the summary statistics of these mean values.

2009). Specifically, a user has unobserved characteristics not captured by data, but these characteristics affect their tweeting behavior, as well their decisions to form links (determining the number of followees) and other users' decisions to follow them (determining the number of followers). Based on the two datasets we have, we devise two methods to approach this endogeneity issue. Furthermore, we are able to cross check our findings using the results we obtain from these two samples.

4.1. Panel Data Approach and Results

Unobserved user characteristics (individual heterogeneity) that are constant over time are an important source of endogeneity in social networks. Personality, as an example, significantly influences tweeting behavior and simultaneously determines peer groups. Besides this, exogenous shocks that affect all Weibo users, *e.g.*, the introduction of a new private messaging tool available to all Weibo users, is another cause of correlations between neighborhood sizes and the error term. To account for these two endogeneity sources, we adopt a panel data method that incorporates both user and time fixed effects.

4.1.1. Empirical Strategy With the longitudinal data structure we assess the effects of the number of followees and those of followers at date $t-1$ on a user's total number of tweets between date $t-1$ and t. A particular data phenomenon is worth noting that, at each cross section, there existed a significant fraction of users who had not had any new tweets since the last period. For instance, $9,788$ of $20,289$ users had no new tweets between November 30, 2011 and December 11, 2011. Figure 4 in Appendix B shows the fractions of users with no new tweets during this period.

This pattern indicates that a propensity to tweet is left censored at zero. A user may have great disincentives to tweet, but all we observe is that she did not tweet during a period of time. To deal with this dichotomy that governs tweeting behavior, we consider two decisions facing each user: the decision whether to tweet; and if a user decides to tweet, how much she would tweet. Thus, we first study the effects of the followee count and the follower count on the decisions to tweet or not.[14] Our main empirical specification is[15]

[14] With skewed distributions of the followee count and the follower count (Table 1 shows that they are both skewed to the right.), we take log-transformations of these variables. Figure 5 in Appendix B displays the distributions of these variables up to September 15, 2011 after log-transformations.

[15] In principle, we could estimate nonlinear models, such as Poisson regressions, to explore the nature of tweets as count data. However, the data suggest that the group of users that did not tweet in each period was very different

$$1\left(\Delta Y_{it} > 0\right) = \beta_1 \cdot \log\left(N^e_{i,t-1}+1\right) + \beta_2 \cdot \log\left(N^r_{i,t-1}+1\right) + \mathbf{X}'_{it}\beta_3 + \mu_i + \nu_t + \epsilon_{it}, \qquad (1)$$

where ΔY_{it} is the user i's total number of tweets (including both tweets and retweets) between date $t-1$ and t; $N^e_{i,t-1}$ and $N^r_{i,t-1}$ are her number of followees and number of followers up to date $t-1$; $\mathbf{X}_{it}$ includes the number of days between date $t-1$ and t; μ_i are user fixed effects, and ν_t the time fixed effects; ϵ_{it} is the idiosyncratic error term.

Given a user decides to tweet (for some t, $\Delta Y_{it} > 0$), we further study the effects of her neighborhood sizes on how much she would tweet in the network. Similarly, we take log-transformation of $(\Delta Y_{it}+1)$. Our main regression equation is

$$\log\left(\Delta Y_{it}+1\right) = \beta_1 \cdot \log\left(N^e_{i,t-1}+1\right) + \beta_2 \cdot \log\left(N^r_{i,t-1}+1\right) + \mathbf{X}'_{it}\beta_3 + \mu_i + \nu_t + \epsilon_{it}, \qquad (2)$$

In both Equation (1) and (2), estimates of β_1 and β_2 will be the effects of followee count and follower count respectively. The simultaneity between tweeting and network sizes causes concerns about serial correlations in error terms. As such, we report the standard errors clustering at the user level in all estimations.

4.1.2. Results The main estimation results are reported in Table 4 and 5. Estimates of our main specifications, Equation (1) and (2), are presented in the fourth column of the corresponding table.[16] The results suggest that, in general, the number of followees in the previous period had negative effects on both whether and how much to tweet, while the lag number of followers had (insignificant) positive effects on both decisions.

More specifically, for the neighborhood size effects on whether to tweet, the results indicate that a 1% increase in the number of followees from last period will lead a user

from those who tweeted. For example, we find that the "silent" users had significantly smaller number of total tweets in stock, fewer followees, and fewer followers. Thus, we explore the two decisions facing each user, which corresponds with the hurdle model that is also standard in the literature on count data (Duan et al. 1983). In addition, given the large numbers of zeros observed in the data, nonlinear models such as a Poisson specification may not work well (Cameron and Trivedi 2005).

[16] For all fixed effects regressions in our panel data analysis, we carry out tests for serial correlations as in Wooldridge (2002). These tests all strongly reject the null hypothesis of no serial correlations. This suggests that we should be careful with inferences. For example, we report the standard errors clustering at user level. In addition, robust standard errors and those from bootstrapping are generally smaller.

Table 4 **Panel Data – Neighborhood Size Effects on Whether to Tweet**

	OLS Results^{a}			
Dep var.: 1(Tweeted)	Spec 1	Spec 2	Spec 3	Spec 4
log (**Lag # Followees**)	0.067***	0.093***	-0.344***	**-0.025*****
	(0.002)	(0.002)	(0.007)	(0.008)
log (**Lag # Followers**)	0.037***	0.037***	-0.005	**0.003**
	(0.001)	(0.001)	(0.004)	(0.004)
# Days Btw Dates	0.006***		0.002***	
	(1.137e-04)		(1.144e-04)	
1(Certified)	-0.050	-0.120***		
	(0.031)	(0.032)		
(Intercept)	0.193***	0.249***	1.789***	0.403***
	(0.008)	(0.007)	(0.027)	(0.031)
User FE	No	No	Yes	Yes
Collection Date FE	No	Yes	No	Yes
Adj. R^2	0.044	0.115	0.335	0.383
Users	20,289	20,289	20,289	20,289
Num. obs.	101,445	101,445	101,445	101,445

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

[a] Robust standard errors are reported. GLS estimates controlling for heteroscedasticity yield similar results.

Table 5 **Panel Data – Neighborhood Size Effects on the Quantity of Tweets**

	OLS Results^{a}			
Dep var.: log (ΔTotal Tweets)	Spec 1	Spec 2	Spec 3	Spec 4
log (**Lag # Followees**)	0.233***	0.335***	-1.435***	**-0.076*****
	(0.007)	(0.007)	(0.027)	(0.025)
log (**Lag # Followers**)	0.113***	0.120***	0.003	**0.005**
	(0.005)	(0.005)	(0.015)	(0.013)
# Days Btw Dates	0.014***		0.002***	
	(3.735e-04)		(3.546e-04)	
1(Certified)	0.051	-0.211		
	(0.146)	(0.146)		
(Intercept)	0.163***	0.155***	6.431***	2.696***
	(0.029)	(0.025)	(0.101)	(0.081)
User FE	No	No	Yes	Yes
Collection Date FE	No	Yes	No	Yes
Adj. R^2	0.047	0.129	0.538	0.628
Users	19,568	19,568	19,568	19,568
Num. obs.	68,954	68,954	68,954	68,954

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

[a] An unbalanced panel containing all cross sections with at least one tweet. GLS estimates controlling for heteroscedasticity yield similar results.

2.5% less likely to tweet during the current period (Table 4). In contrast, an increase in the lag number of followers increases the probability of tweeting (Table 4). The neighborhood size effects on the quantity of tweets exhibit the same pattern. Specifically, a 1% increase in the number of followees reduces, on average, 0.08% of tweets up to the next observation date (Table 5). However, a 1% increase in the number of followers tends to increase the number of tweets by 0.5 percentage point (Table 5). Note that if we do not include user or time fixed effects, both the number of followees and the number of followers are positively related to both decisions whether and how much to tweet (by comparing Spec 1 with Spec 4 in both Table 4 and 5). This illustrates the potential bias due to the endogeneity of an individual's networks or peer groups.

4.2. Cross-Sectional Data and Results

As briefly mentioned, some concerns on the endogeneity of group sizes, such as the correlation between the endogenous variables and user unobserved characteristics that are constant over time, are alleviated by including user and time fixed effects. However, other types of correlations might still exist between the error term and endogenous variables. As an illustration, the event of an individual's marriage during our study period might have long lasting effects on both network formation decisions and tweeting behavior but it cannot be captured by either user or time fixed effects. Thus we need more evidence from our cross-sectional data that allows us to address the endogeneity problem due to the user-time-specific heterogeneity and reverse causality.

Our cross-sectional data capture a user's network of followers and followers as well as the higher order of networks. Exploiting this feature, we combine a strategic network formation process into our tweeting equation and use the excluded variables in the network formation equation as instrumental variables. The following, section 4.2.1 lays down the network formation model, section 4.2.2 discusses the assumptions necessary for instrument validity, and section 4.2.3 presents our empirical strategies and findings.

4.2.1. The Model The model prevalent in peer effects literature is the linear-in-means model (Bramoullé, Djebbari, and Fortin 2009, Goldsmith-Pinkham and Imbens 2013, Peng et al. 2016). We combine the model adapted to our situation with a strategic network formation process to develop instrumental variables. Consider first that a user's tweeting equation on Tencent Weibo has the linear-in-means form

$$\log(Y_i) = \beta_0 + \beta_1 \cdot \log(N_i^e) + \beta_2 \cdot \log(N_i^r) + \mathbf{X}_i' \beta_3 + \beta_{\bar{Y}} \cdot \bar{Y}_{(i)} + \bar{\mathbf{X}}_i' \beta_{\bar{\mathbf{X}}} + \epsilon_i, \tag{3}$$

where Y_i is user i's number of tweets; N_i^e and N_i^r are her number of followees and followers (who are not simultaneously followees) respectively; $\mathbf{X}_i$ is a vector of her observed characteristics including age and gender; $\bar{Y}_{(i)}$ is the average number of tweets by i's all followees and followers, and $\bar{\mathbf{X}}_i'$ are the means of observed characteristics of all her followees and followers. Particularly, the error term ϵ_i captures all remaining factors that affect tweeting, such as user i's unobserved characteristics and idiosyncratic shocks.

Let $\mathbf{X} = [\mathbf{X}_i]$ be the observed characteristics of all users and ϵ the vector of all user error terms that capture their unobservables. We assume that they are independent. Specifically,

Assumption 1. The observed characteristics are orthogonal to unobserved characteristics, i.e., $\mathbf{X} \perp \epsilon$.

A major concern of causality arises because of the endogeneity of an individual's peer groups, which are followees and followers in our context. Under our current context, two issues are most relevant: the simultaneity problem and unobserved heterogeneity. For the first aspect, a user's tweeting behavior is not only affected by the size of her neighborhood but also determines the number of other users who form connections with her. For the other, unobservables affect her tweeting behavior, as well as network formation decisions. Since a user's peer groups are endogenous, in Equation (3) N_i^e, N_i^r, $\bar{Y}_{(i)}$, and $\bar{\mathbf{X}}_i$ are all endogenous variables. Econometrically, these endogenous variables are correlated with the error term ϵ_i.

We propose to use a set of instrumental variables to approach the endogeneity issue. Specifically, the IVs are the average observed characteristics, including age and gender, of a user's second-order followees and second-order followers who do not belong to her immediate neighbors.

We obtain the ideas for our IV strategy from the network formation process. Suppose user i is determining whether to follow another user j. Many factors may influence her connection decisions. First, since her followees' tweets are her major source of information on Tencent Weibo, user i cares about what and how much j tweets. She may also examine whether j has similar (both observed and unobserved) characteristics. In social networks, a user tends to follow others who are similar (the phenomenon of homophily, such as McPherson, Smith-Lovin, and Cook (2001)). On top of these, idiosyncratic shocks may also affect user i's utility of following j. For example, the occurrence of a natural disaster makes a user more likely to follow news reporters or charity organizations. We thus formalize i's utility of following j as: for any $j \neq i$,

$$u_{ij} = f\left(Y_i, Y_j; \mathbf{X}_i, \mathbf{X}_j; \epsilon_i, \epsilon_j\right) + \eta_{ij}, \tag{4}$$

where $f(\cdot)$ is a function of the quantity of tweets and all characteristics for both i and j; and η_{ij} is the idiosyncratic error term representing shocks that are i-j pair specific.

Given i's utility of following j, her decision to follow j obeys a simple rule that

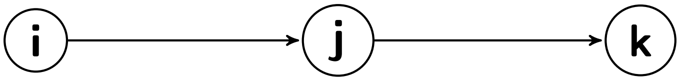
$$\begin{cases} i \text{ follows } j, & \text{if } u_{ij} \geq 0; \\ i \text{ does not follow } j, & \text{if } u_{ij} < 0. \end{cases}$$

As a consequence, these network formation processes lead to variations in peer effect variables in Equation (3), including N_i^e, N_i^r, $\bar{Y}_{(i)}$, and $\bar{\mathbf{X}}_i$. This is the basis for our IV strategy: variables excluded in the question (3), for example a user's second-order neighbors' characteristics, changes her tweeting incentives only by changing her network size.

4.2.2. Validity of Instruments Valid instruments are excluded variables in the tweeting equation and are correlated with the group sizes and uncorrelated with the unobservables. Based on the model in the last section, we discuss the validity of our instruments. Figure 1 is a parsimonious social network with only three users. Our discussions will be illustrated by this example. Specifically, suppose there are three Weibo users, i, j, and k, and two edges (following relations) connecting them. The arrows in Figure 1 show the following directions, e.g., user i is a follower of user j and thus j is i's followee.

Relevance: The first set of instruments contains all the average characteristics of second-order followees. In Figure 1, user k's characteristics will be used as instruments for the number of followees of user i. To see how they are relevant, they enter user j's tweeting equation (Equation (3) for j) since k is j's immediate followee. This implies that the instruments have direct effects on j's actions, tweeting, which in turn determines i's utility of following j, since user j's tweeting enters Equation (4) for i. In other words, variations in our IVs lead to changes in a user's utility of forming connections, indirectly by changing the first-order followees' tweeting activities. This process governs the number of followees an individual might have (N_i^e). Similarly, the average observed characteristics of second-order followers, our second set of instruments, are correlated with the number of followers since they enter immediate followers' tweeting equations, which in turn determine the connection decisions of immediate followers.

Figure 1 A Simple Network with Three Users and Two Edges



Correlations between the average observed characteristics of a user's second-order neighbors and the tweets and characteristics of her immediate neighbors ($\bar{Y}_{(i)}$ and $\bar{\mathbf{X}}_i$ in Equation (3)) are easier to establish, since they are mutually first-order neighbors.

Exclusion restriction: Clearly, the validity of our IV strategy requires a certain degree of shortsightedness when a user is tweeting. Specifically, we require the assumption that individuals are not perfectly forward-looking, in the sense that when tweeting they are only affected by the characteristics of immediate followees and followers but not of neighbors' neighbors (implicitly assumed in Equation (3)). Sociology literature (Murthy 2012) argues that Twitter users "are not always consciously aware that their tweets have the potential to travel further (than immediate followers)." As a result, the average observed characteristics of second-order neighbors do not enter a user's tweeting equation—Equation (3).

In addition, we also assume shortsightedness when a user is making network formation decisions. As implicitly assumed in Equation (4), the characteristics of neighbors' neighbors are excluded from the utility function. By using similar arguments as above, the information about the indirect connections does not affect connection decisions.

Exogeneity: It is left to show that our instruments are exogenous, i.e., the average observed characteristics of second-order neighbors is uncorrelated with the error term ϵ_i in Equation (3). It turns out that Assumption 1 is a necessary condition for exogeneity. To motivate this assumption, consider ϵ being the individual idiosyncratic component influencing her network formation utilities. It captures all remaining effects that cannot be explained by an individual's observed characteristics, $\mathbf{X}$. It is often assumed in the literature (Xiao 2010) that the observed and unobserved components capture separate effects on outcomes, and thus are orthogonal to each other.[17]

[17] Relaxing Assumption 1 correlates our instruments with a user's unobservables, because users' unobserved characteristics are first correlated with their immediate neighbors' characteristics (both observed and unobserved), and then in turn correlated with our instrument. Take Figure 1 as an example; we first consider $\mathbf{X}_k$ as instruments for i's number of followees. Suppose Assumption 1 does not hold and that $\mathbf{X}_k$ will be first correlated with $\mathbf{X}_j$ and ϵ_j governed by j's decision to follow k (Equation (4) for j). Then as ϵ_j is correlated with ϵ_i (homophily), our instruments $\mathbf{X}_k$ are indirectly correlated with ϵ_j.

4.2.3. Empirical Strategy and Results Our main empirical specification for cross-sectional analyses is based on the linear-in-means model, Equation (3). Again, estimates of β_1 and β_2 will be the effect of peer group sizes on a user's quantity of tweets as of the data collection date with an elasticity interpretation.

We use the average observed characteristics of second-order followees and those of second-order followers (but not immediate neighbors) as instruments for all the endogenous variables, $\log(N_i^e)$, $\log(N_i^r)$, $\bar{Y}_{(i)}$, and $\bar{\mathbf{X}}_i$. More specifically, the observed characteristics contain the user age, gender dummy for female, dummies for missing values of age and gender separately, and a dummy for age outliers.[18] Then for each user, we calculate the mean value of each characteristic of her second-order followees and second-order followers correspondingly. For the average tweets and characteristics of an individual's reference groups ($\bar{Y}_{(i)}$ and $\bar{\mathbf{X}}_i$), we calculate these (endogenous) variables by stacking together her immediate followees and followers.

We report our main results in Table 6. In this table, Spec 1 to 3 report the OLS estimates of Equation (3) with the followees only, the followers only, and both followees and followers. We report the two-stage least squares (2SLS) estimates with IVs in the last column. Results of all first-stage regressions are reported in Table 14 in Appendix D.[19] We can see that using IVs changes the sign for the followee effect and magnifies the follower effect. Results in the last column are consistent with our findings from the panel data analysis, that is, we have negative effect of the number of followees and positive effect of the follower count. However, the magnitudes are different. Particularly, our estimates suggest that a

[18] We define users with year of birth being before 1940 or after 2000 as age outliers. We include this dummy variable to control for cases of "fake" year of birth, a variable that is reported by users.

[19] One may worry that excluding the average characteristics of immediate neighbors in our IV may lead to the issue of weak instruments. First, however, the results from the first-stage regressions (F-statistics in particular) suggest that we may not have such problems. More importantly, from the early discussions on the validity of instruments, our IVs are not likely to be weakly correlated with the endogenous variables.

Table 6 Cross-Sectional Data – Main Estimation Results

	OLS Results[a]			**2SLS Results**
Dep var.: log (Total Tweets)	Spec 1	Spec 2	Spec 3[b]	
log (# **Followees**)	0.358***		0.272***	**-0.194***
	(0.002)		(0.002)	(0.105)
log (# **Followers**)		0.480***	0.363***	**1.306*****
		(0.003)	(0.003)	(0.186)
Characteristics of All Neighbors				
Avg. Total Tweets	0.346e-03***	3.226e-04***	3.251e-04***	-1.164e-04
	(0.590e-07)	(0.595e-07)	(0.558e-07)	(3.161e-04)
Avg. Age	-0.018***	0.002***	-0.010***	-0.035***
	(3.237e-04)	(3.152e-04)	(0.307e-03)	(0.004)
Avg. 1(Missing Year Birth)	-0.340***	0.752***	0.115***	-9.166**
	(0.036)	(0.038)	(0.036)	(4.021)
Avg. 1(Female)	0.010	0.168***	0.123***	-0.540
	(0.012)	(0.012)	(0.012)	(0.604)
Avg. 1(Missing Gender)	-0.538***	0.827***	-0.033	13.033**
	(0.041)	(0.043)	(0.040)	(6.177)
Avg. 1(Age Outliers)	-0.263***	0.371***	-0.004	-0.952
	(0.019)	(0.020)	(0.019)	(0.845)
User Characteristics				
Age	-0.001***	-0.002***	-0.002***	-0.002***
	(1.199e-04)	(1.212e-04)	(0.116e-03)	(0.424e-03)
1(Missing Year Birth)	-0.236***	-0.259***	-0.204***	-0.110***
	(0.018)	(0.019)	(0.018)	(0.040)
1(Female)	0.105***	0.090***	0.120***	0.245***
	(0.004)	(0.004)	(0.004)	(0.037)
1(Missing Gender)	-0.149***	-0.188***	-0.148***	-0.146***
	(0.026)	(0.026)	(0.025)	(0.041)
1(Age Outliers)	0.026***	-0.045***	-0.011*	-0.135***
	(0.007)	(0.007)	(0.007)	(0.022)
(Intercept)	4.206***	4.071***	3.869***	5.776***
	(0.013)	(0.014)	(0.013)	(0.568)
Instruments	No	No	No	Yes
Adj. R^2	0.132	0.131	0.190	-
Num. obs.	402,470	402,470	402,470	402,470

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

[a] We present OLS estimates of various specifications of Equation 3 in columns 2 to 4. We also report the two-stage least squares (2SLS) estimates in the last column using the average characteristics of second-order followees and followers as instruments for the endogenous variables. In all regressions, the dependent variable is log (# Total Tweets + 1) since there exist observations with no tweets at all. A GMM method considering heteroscedasticity yields similar estimates and statistical inferences.

[b] Comparing estimates in Spec 3 with those from 2SLS estimation, we find that ignoring the endogeneity overestimates the followee size effect, while underestimating the effect of followers.

1% increase in the number of followees will reduce a user's quantity of total tweets by about 0.2%. On the other hand, a 1% increase in the follower count raises the number of total tweets by more than 1.3%. Both effects are statistically significant.

Interestingly, our results suggest that ignoring the endogeneity of a user's peer groups (Spec 3 in Table 6) will overestimate the effect of followee count, while underestimating the effect of the number of followers. We attribute these bias corrections to the correlations between the endogenous group sizes and the error term. For example, a person's "chattiness" is an important factor leading to more tweeting but omitted from **X**'s (therefore picked up by the error term). A more chatty user wants more information from her followees, which leads to a positive correlation between the error term and the endogenous followee count. On the other hand, a chatty user may be more cautious about expressing herself with a large audience, which implies a negative relationship between the follower count and the error term.

Our results also show some interesting patterns in demographic groups. Younger Weibo users tweet more on average, as the estimate in Table 6 suggests a negative relationship between the total number of tweets and age. More interesting, a user with a group of younger friends also tweets more controlling for all other characteristics. We also observe that female users tend to tweet more than their male fellows. The estimate in Table 6 suggests that female users tweet about 24.5% more than male users.

4.3. Robustness Checks

To ensure the robustness of our empirical results, we address several concerns about our empirical specifications and the samples we use in this section. We find our main results are robust to changes in the formulation of an individual's tweeting, the exclusion of "inactive" users, and alternative specifications.

4.3.1. Panel Data: Unbalanced panel: As mentioned in the data section, the original sample of our panel data consists of $29,956$ users. For some of them, observations are missing at some dates. This missingness is due to the data provider failing to collect those records. Some may suspect that these "incomplete" users may carry extra information that is not captured by our main sample, potentially attenuating our results. We conduct our

Table 7 Robustness Check (Panel) I, II, and III – Alternative Samples

	OLS Results						
	(I)[a]: Unbalanced Panel		**(II)**[b]: Tweets/Retweets			**(III)**[c]: New neighbors	
Dep var.:	1(Twtd)	log (ΔTT)[d]	1(Twtd)	log (ΔT)	log (ΔRT)	1 (Twtd)	log (ΔTT)
log (**Lag # Flwees**)	**-0.165*****	**-0.371*****	**-0.034*****	**-0.079*****	**-0.004*****	**-0.023*****	**-0.070*****
	(0.005)	(0.021)	(0.008)	(0.025)	(0.001)	(0.008)	(0.025)
log (**Lag # Flwers**)	**0.007***	**0.004**	**0.005**	**0.009**	**0.002*****	**0.267e-04**	**0.013**
	(0.003)	(0.011)	(0.004)	(0.013)	(0.001)	(0.004)	(0.014)
log (New # Flwees)						0.078***	0.238***
						(0.016)	(0.076)
log (New # Flwers)						0.082***	0.270*
						(0.028)	(0.160)
(Intercept)	1.359***	3.679***	0.982***	2.495***	5.849***	-0.232	-1.183
	(0.017)	(0.069)	(0.025)	(0.080)	(0.003)	(0.229)	(1.250)
User FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adj. R^2	0.378	0.602	0.396	0.594	0.519	0.383	0.628
Users	29,956	28,426	20,289	19,059	19,059	20,289	19,568
Num. obs.	139,007	92,518	101,445	64,292	64,292	101,445	68,954

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

[a] In this panel, we present the estimation results of Equation 1 and 2 using the unbalanced panel.

[b] Estimation results of Equation 1 and 2 based on the number of tweets/retweets only.

[c] Estimates of Equation 5 are reported in this panel.

[d] In this table, "T" stands for tweets; "RT" is short for retweets; and "TT" is total tweets.

first robustness check with the unbalanced panel and report the estimates of Equation 1 and 2 in the panel I of Table 7. The results are qualitatively identical to our main results.

Tweets/retweets only: In the main specification and all robustness checks, the dependent variables are based on a user's number of total tweets that is the sum of tweets and retweets. As the behavior of writing tweets may be quite different from that of retweeting others' tweets, we conduct our second robustness check by constructing the dependent variables based on tweets only, or retweets only. We report the estimation results in Table 7 and again find qualitatively similar effects as with our main results.

New neighbors: A third concern is about the empirical specifications used in our panel data analyses. One may argue that not only the number of followees and followers in the last period but also the number of new followees and new followers between date $t-1$ and t affect a user's tweeting behavior. We thus add two additional controls to Equation 1 and 2 and estimate the following specifications:

$$\mathrm{DV} = \beta_1 \cdot \log\left(N^e_{i,t-1}+1\right) + \beta_2 \cdot \log\left(N^r_{i,t-1}+1\right) + \beta_3 \cdot \log\left(\Delta N^e_{it}\right) + \beta_4 \cdot \log\left(\Delta N^r_{it}\right) + \mathbf{X}'_{it}\beta_5 + \mu_i + \nu_t + \epsilon_{it}, \quad (5)$$

where DV can be the tweeting indicator $1(\Delta y_{it} > 0)$ and new tweet count $\log(\Delta y_{it}+1)$; ΔN^e_{it} and ΔN^r_{it} are the change in the number of followees and followers respectively. Estimates are reported in Table 7. Our main findings on the neighborhood size effects are robust to this specification concern. It is also worth noting that the effects of both the number of new followees and that of new followers on either the propensity or the intensity of tweeting are positive (Table 7).

4.3.2. Cross-Sectional Sample: Tweets/retweets only: We perform similar robustness checks for our analyses of the cross-sectional sample. Specifically, we also examine whether using tweets (or retweets) instead of the total number of tweets and retweets would alter our results, a situation which arises because of concerns about different behaviors underlying tweeting and retweeting. We thus calculate the dependent variables in Equation (3) based on the number of tweets (or retweets) and report the 2SLS estimates in the first two columns of Table 8. The effects of the followee count and follower count are shown to be not significantly different from those in our main results. In particular, we observe even stronger effects of both the followee and follower counts on the quantity of retweets.

Inactive users: Another robustness check for the cross-sectional analyses is to examine the effects of "no tweeting." We find that, in the cross-sectional sample, 992 (out of 402,470) users did not have any tweet as of the data collection date. We check whether the behavior of "active" users is significantly different by excluding these inactive users. We report the 2SLS estimates in the third column of Table 8. The peer group size effects are qualitatively similar to those from our main sample.

"Outliers": The sample of users in our cross-sectional analyses may be another concern. Users with large numbers of followers are usually "VIPs" or celebrities in the offline world.

Table 8 Cross-Sectional Data – Robustness Checks

	2SLS Results[a]			
	Alternative Specs		**Alternative Samples**:	
			No "Inactive" Users[b]	No "Outliers"[c]
	(1)	(2)	(3)	(4)
Dep var.:	log (Tweets)	log (Retweets)	log (Total Tweets)	log (Total Tweets)
log (# **Followees**)	**-0.116**	**-0.749*****	**-0.233****	**-0.161***
	(0.095)	(0.253)	(0.108)	(0.097)
log (# **Followers**)	**0.970*****	**2.566*****	**1.359*****	**1.232*****
	(0.173)	(0.395)	(0.187)	(0.170)
Characteristics of All Neighbors				
Avg. Total Tweets			-1.737e-04	-1.796e-04
			(3.318e-04)	(3.019e-04)
Avg. Tweets	-2.433e-04			
	(3.436e-04)			
Avg. Retweets		0.011***		
		(0.004)		
Avg. Age	-0.042***	0.021**	-0.034***	-0.036***
	(0.004)	(0.010)	(0.004)	(0.004)
Avg. 1(Missing Year Birth)	-7.472**	17.353	-8.893**	-8.669**
	(3.312)	(14.149)	(4.110)	(3.820)
Avg. 1(Female)	-1.070*	2.167	-0.672	-0.772
	(0.557)	(1.372)	(0.607)	(0.587)
Avg. 1(Missing Gender)	7.687	26.590**	12.782**	10.313*
	(5.766)	(12.946)	(6.184)	(5.721)
Avg. 1(Age Outlier)	-1.588**	8.572***	-0.644	-1.008
	(0.699)	(3.035)	(0.892)	(0.803)
Control variables.	Yes	Yes	Yes	Yes
Instruments	Yes	Yes	Yes	Yes
Num. obs.	402,470	402,470	401,478	401,640

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

[a] This table presents four robustness checks of the cross-sectional analyses. Two-stage least squares estimates of our main specification, Equation (3), are shown here. GMM estimations yield qualitatively similar results.

[b] In this robustness check, we delete the "inactive" users who had no tweets by the data collection date.

[c] We eliminate "outliers" with more than $1,000$ followers in this robustness check.

Their incentive to use the online networking tool may be quite different from an "average" person. For instance, a politician in a campaign may tweet to attract more followers and then potentially more supporters in offline elections. We examine whether our main results are mainly driven by these "outliers" by performing another robustness check. We check whether the main results still hold if these users are excluded. Specifically, we eliminate the users with more than $1,000$ followers,[20] the set of which contains 830 users. Results

[20] We also consider other thresholds such as 100, 200, and 500. The 2SLS estimates of our key variables are qualitatively similar to our main results.

are reported in column (4) of Table 8. Both the effect of followees and that of followers are slightly smaller in absolute values, which implies that the tweeting of these VIP users is more "elastic" with respect to neighborhood sizes.

5. Discussions and Implications

5.1. Negative Effects of the Followee Count

Although our results from longitudinal and cross-sectional analysis differ in magnitude, the qualitative results are the same: the number of followers has positive effects on a user's decision to tweet, while the effects of the followee count are negative. Previous studies have focused almost exclusively on an individual's followers (Zhang and Zhu 2011, Toubia and Stephen 2013) and unanimously find positive impacts. One of our contributions is to divide the peer influence into two parts: in-network (followees) and out-network (followers). We consider theoretical explanations of the negative followee effect from several aspects below.

First, as we have already argued, a social media user has conflicting incentives to contribute: free riding versus maximizing social benefits. The negative impact of the followee count shows that the user has the incentive to free ride the contributions of others, in particular followees. Second, and closely related to the first point, social media users are simultaneously information consumer and producer. When the acquisition of information is not free, social communication aggravates an individual's incentives to consume the information generated by others (free riding), than to aggregate and produce information themselves (Han and Yang 2013). Last but not least, online attention is a scarce resource that online platforms compete for (Iyer and Katona 2016). A user distributes his or her attention on an online platform between consuming (reading tweets by followees) and producing (tweeting). With a larger set of followees generating more tweets, the tradeoff between consumption and production apparently favors the former.

5.2. Comparing Panel with Cross-Sectional Analysis

Comparing results from the panel data analysis and those from the cross-sectional data, we find that the directions of group size effects are the same while the magnitudes differ. The inconsistency in magnitudes is partly driven by the relatively small sample size of our panel data. In our panel sample, we have records of $20,289$ users, which is fairly small compared with the Weibo population of over 500 million registered users. In contrast, we have over 2 million users in our cross-sectional data. The sample comprises of a non-ignorable subsample of the Weibo population. More important, as we argue in Section 4.1.2, extra correlations exist between the endogenous variables and the error term even after we incorporate both user and time fixed effects. Because of these extra concerns, our results indicate that the panel data analysis overestimates the effect of the followee count while underestimating that of the follower count. This is consistent with our discussions of bias corrections in cross-sectional analysis (Section 4.2.3) and supported by the data.[21]

In addition, the empirical specifications in panel data analysis and cross-sectional analysis are different. In particular, the network structure in cross-sectional analysis allows us to incorporate the peer effects terms ($\bar{Y}_{(i)}$ and $\bar{\mathbf{X}}_i$ in Equation 4.2), which are absent in the panel analysis since we do not observe the identities of an individual's neighbors. The literature (Manski 1993, Goldsmith-Pinkham and Imbens 2013) establishes important effects of peer activities and characteristics on behavior, which should not be omitted. For these reasons, the results of the cross-sectional analysis are more trustworthy. Our counterfactual exercises later will be based on these results.

[21] These findings provide further evidence that the error term is positively correlated with the number of followees while negatively correlated with the follower count. As an example, suppose the level of a user's dependence on the platform (varying across users and over time) enters the error term. This factor predicts a positive relationship between the number of followees and the error term, since more dependent users normally require more information from the platform. In contrast, such users are generally more cautious about their platform images and therefore more cautious about tweeting with a larger population of followers.

5.3. The Moderating Role of User Characteristics

Are the average effects we have identified so far the same across different groups of users? We examine two prominent user characteristics that can potentially moderate the main effects, namely the "activeness" of a user's followers and gender.

We propose to use the average number of tweets plus retweets as a measure of an individual's followers' activeness on site. We differentiate the sampled users by this measure and notice that one third of the sample provides fewer than 146 total tweets, one third more than 357, and the rest in between. We carry out subsample analysis, and the results are reported in Table 9. Interestingly, we notice that the effect of the follower count increases with follower activeness. In other words, as followers are more active in disseminating the content (instead of merely clicking on the notification of tweets received), it offers more incentives for the content generator (the focal user) and thus she tweets more.

As one of the important contextual factors, gender has been shown to play a moderating role in Internet users' online activities (Ghose and Han 2011). We examine whether there are any differential effects across gender groups. We separate our main sample into groups of female users and male users and report the results in Table 9. We find that the size of peer groups has a larger influence on female users, regardless of the peer groups. More specifically, both the negative followee effect and the positive follower effect are larger in absolute values for female users.

5.4. Managerial Implications

From a practical point of view, especially the platform's purpose, Tencent Weibo wants to increase user "activeness"—tweeting more. A significant fraction of the platform's profits come from the click-through rate of the advertisements (Xiang 2012). Higher level of user participation leads to more click-throughs of the advertisement. Our results have implications for platform managers hoping to promote more active contributions.

Table 9 Cross-Sectional Data – Moderating Effects

	2SLS Results[a]				
	Avg. total tweets by followers			**Gender**	
Dep var.: log (Total tweets)	≤ 146	$(146, 357]$	> 357	Female	Male
log (# **Followees**)	**0.007**	**0.033**	**-0.843**	**-0.331***	**-0.040**
	(0.270)	(0.113)	(0.769)	(0.200)	(0.146)
log (# **Followers**)	**0.662**	**1.106*****	**2.074***	**1.423*****	**0.999*****
	(1.605)	(0.290)	(1.102)	(0.341)	(0.171)
Characteristics of All Neighbors					
Avg. Total Tweets	-0.001	0.001**	0.001	0.001	-0.001
	(0.002)	(0.472e-03)	(0.001)	(0.001)	(0.001)
Avg. Age	-0.037**	-0.036***	-0.024	-0.022*	-0.014*
	(0.015)	(0.008)	(0.027)	(0.012)	(0.009)
Avg. 1(Missing Year Birth)	8.094	-5.423	-14.538	3.237	-2.088
	(13.269)	(5.240)	(21.317)	(8.114)	(11.870)
Avg. 1(Female)	-5.303**	0.399	3.062	-0.378	1.403**
	(2.177)	(1.054)	(2.506)	(0.832)	(0.671)
Avg. 1(Missing Gender)	-26.790**	11.239	61.636*	11.018	-4.156
	(10.901)	(9.160)	(33.869)	(8.179)	(8.938)
Avg. 1(Age Outlier)	-0.767	-3.137***	4.859	3.101	-0.515
	(1.942)	(1.205)	(7.407)	(2.784)	(1.358)
Control variables	Yes	Yes	Yes	Yes	Yes
Instruments	Yes	Yes	Yes	Yes	Yes
Num. obs.	132,667	136,999	132,804	205,583	196,887

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

[a] This table presents two sets of subsample analysis. In the first set, we categorize the sampled users by the activeness of followers, measured by the average total tweets among all followers. The results are reported in the first three columns. In the second set of subsample analysis, we differentiate the users by their gender. Results are reported in the last two columns.

Platforms with a similar network structure can potentially benefit from strategies to promote more connections among users, but the asymmetric effects of the follower count and the followee count act as a double-edged sword. To assess the impact of new connections on tweeting, we first study the marginal effects of the follower count on a user's tweets. Based on our estimates, with the majority of users having one follower $(180,225$ out of $402,470$ in the cross-sectional sample), hypothetically adding one additional follower will more than double their quantity of tweets (130.6% increase) holding everything else constant. Similarly, a user with 2 followers (the median of follower count distribution) will tweet 65.3% more with 1 additional follower. We visualize the estimated marginal effects of the followee count and the follower count in Appendix E.

Although the positive effect of the follower count on a user's tweeting is both economically and statistically significant, the effect of additional links on overall tweeting in the network is still ambiguous. With one extra link in the network, a user has one more follower while the other has an additional followee. The negative effect of the followee count on one's contributions masks the effect on overall contributions in the network. To determine the overall effect, we conduct a series of simulations.

5.4.1. Simulated Effects of Random New Links We use the total contributions, more specifically the total number of tweets by all users, as a measure of "activeness." We also use other statistics about the distribution of tweets, such as mean and median, for similar comparisons. In our counterfactual exercises, we first randomly generate new links to the existing network. For each draw of a simulated network, we are able to trace out the effect of changes in neighborhood sizes (followee and follower counts) on the total tweets for each user. To see this, we denote $\tilde{Y}_i$ the number of total tweets the user i would tweet in the simulated network. Suppose the number of followees and followers are $N_i^e + \Delta N_i^e$ and $N_i^r + \Delta N_i^r$ respectively in the simulated network. By our main specification, $\tilde{Y}_i$ will be,

$$\log\left(\tilde{Y}_i\right) = \beta_0 + \beta_1 \cdot \log\left(N_i^e + \Delta N_i^e\right) + \beta_2 \cdot \log\left(N_i^r + \Delta N_i^r\right) + \beta_{\bar{Y}} \cdot \bar{Y}_{(i)} + \bar{\mathbf{X}}_i' \beta_{\bar{\mathbf{X}}} + \mathbf{X}_i' \beta_3 + \epsilon_i. \quad (6)$$

By comparing Equation (3) with (6), we show that the change in tweets, $\tilde{Y}_i - Y_i$, is

$$\tilde{Y}_i - Y_i = Y_i \cdot \frac{\left(N_i^e + \Delta N_i^e\right)^{\beta_1} \cdot \left(N_i^r + \Delta N_i^r\right)^{\beta_2}}{\left(N_i^e\right)^{\beta_1} \cdot \left(N_i^r\right)^{\beta_2}} - Y_i. \quad (7)$$

Then based on the 2SLS estimates $\hat{\beta}_1$ and $\hat{\beta}_2$, we can calculate the estimated change in total tweets for each user i. This allows us to back out the empirical distribution of $\tilde{Y}_i - Y_i$, and examine whether random new connections between users lead to higher levels of contributions in the network.

We draw three different simulated networks, with 0.5 million, 1 million, and 2.5 million new links respectively. Given the number of new links, different draws of simulated networks yield almost overlapping distributions of the change in tweets. Therefore in Figure 2 we depict, for each given number of extra links, the empirical density function of $\tilde{Y}_i - Y_i$ from just one simulated network (no need to show overlapping distributions from different draws). We first notice that with more simulated links the distribution has fatter tails. Unlike what the figures might suggest, the distributions are in fact highly skewed to the right. For example, with 2.5 million new links, the median is 129.82 while the mean is 343.62. Table 10 summarizes the quantiles and the total changes in tweets, a column corresponding to a random draw. We observe that with a greater number of simulated links, the total contribution measured by the total quantity of tweets increases significantly. Particularly, with less than a 1% increase in the number of links (0.5 million), the total tweets increase by about 25.22%. More dramatically, the 2.5 million new links (about 5% increase from the observed network) more than double the total number of tweets (increase by 133.75%). These dramatic effects are mainly driven by users with a very small number of followers. We notice that, for all three simulations, more than 97% of the total change in tweets are attributed to the contributions by the subgroup of users with fewer than 5 followers. It is also clear, by comparing the quantiles, that the overall contributions are growing with more links.

5.4.2. Simulated Effects of Targeted New Links Furthermore, the platform favors recommendations to the users with a small number of followers. For example, a more effective way is to recommend users to follow those who have fewer followers (targeted recommendations) instead of popular ones who have established a large number of followers. Since the marginal effect of extra followers is larger and dominant at a small number of followers (relative to the effect at a large number of followers),[22] this method will generate even more contributions than purely random recommendations.

[22] In Appendix F we report the fractions of new tweets generated by users with different number of followers by fixing the number of followees.

Figure 2 **Distributions of the Change in Tweets $\tilde{Y}_i - Y_i$ from Three Simulated Networks**

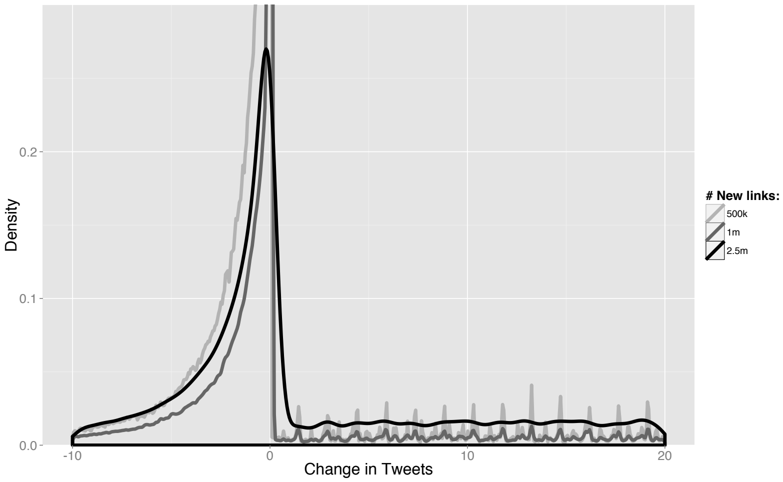


Table 10 **Comparing the Sample Quantiles of the Distributions of Changes in Tweets**

	# New Links		
Quantiles	**0.5 Million**[a]	**1 Million**	**2.5 Million**
10%	-1.565	-2.327	-1.042
20%	-0.119	-0.470	9.153
30%	0.000	0.000	40.780
40%	0.000	0.000	79.549
50%	0.000	4.187	129.817
60%	0.000	42.683	199.957
70%	0.000	95.729	304.668
80%	64.972	182.758	480.711
90%	194.004	370.051	866.620
Sum of Changes in Y_i	2.608e+07	5.300e+07	1.383e+08
Sum of Y_i	1.034e+08	1.034e+08	1.034e+08
Percentage Change[b]	25.222%	51.256%	133.749%

[a] In this table, we present the sample quantiles of the distributions of changes in total tweets, $\tilde{Y}_i - Y_i$, from three different simulations. Each simulation has different number of new links.

[b] The variable calculates the ratio of the sum of changes in Y_i relative to the sum of Y_i, *i.e.*, the percentage change of total simulated tweets relative to the observed total tweets.

Given the follower effects are dominant for occasional users, we then study whether targeting these users will promote a higher level of activeness in the social network. We conduct a set of new counterfactuals in which we generate a certain number of new links that follow the targeted subsample of users, those with fewer followers. Specifically, we run the following simulations and compare:

(a) *Original simulation*: 1 million new links among the full sample of $1,392,873$ users;

(b) *Targeted simulation I*: 1 million new followers to the users with less than 5 followers;

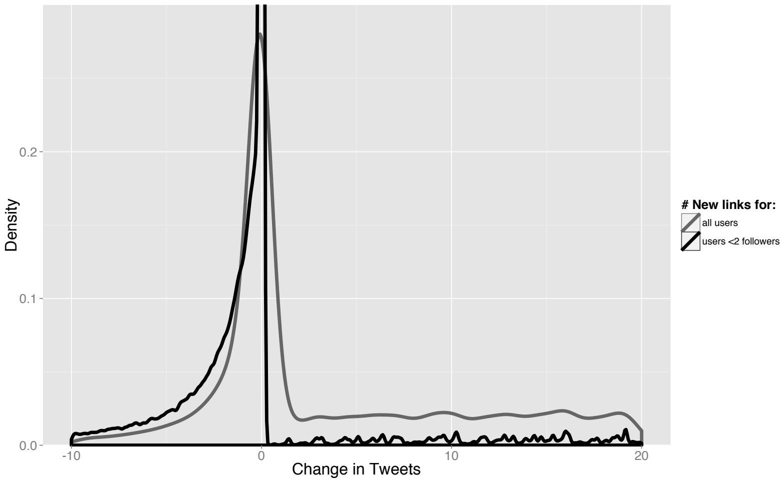
(c) *Targeted simulation II*: 1 million new followers to the sample used in our main empirical specification (totally $402,470$ users);

(d) *Targeted simulation III*: 1 million new followers to the users with fewer than 2 followers in the sample used in our main empirical specification.

We study two sets of comparisons: comparing simulation (a) to (b) and simulation (c) to (d). We find, in both comparisons, that the simulations targeting the users with fewer followers generate more tweets. Student t-tests suggest that users in simulation (b) (or (d) in the other set of comparison) add more tweets than users in (a) (or (c) correspondingly) at any usual statistical significance level. In particular, the quantity of new tweets generated in simulation (d) is about 23% more than that from simulation (c). Not surprisingly, the distribution of the change in tweets from simulation (b) (or (d) in the other comparison) has much longer tail than that from (a) (likewise (c)). Figure 3 displays the two distributions from simulation (c) and (d). These findings confirm our conjecture that user recommendations targeting less popular users will promote even higher level of contributions than purely random recommendations.

From the practitioners' point of view, our results not only recommend strategies the platform can use to foster higher levels of contributions and thus to improve its long-term viability but also have implications for commercial users on social media. On this particular platform, Tencent Weibo, we find that the positive follower effect is much stronger than the negative followee effect, so such marketing efforts can be effective. On another platform, however, similar marketing strategies may be unsuccessful. We hope that our results

Figure 3 **Distributions of the Change in Tweets $\tilde{Y}_i - Y_i$ from Simulations (c) and (d)**



encourage the platforms to consider this tradeoff, and take into account a social media user's conflicting incentives in generating and disseminating information in the network.

It is common practice that firms take advantage of the advertising slots embedded in social media timelines (Miller and Tucker 2013). Facebook's "Suggested Posts" and Twitter's sponsored tweets are two examples. One of the features, which is highly salient in these posts, is to attract more followers to the commercial pages. According to our results, these campaigns can have unintended consequences in that the efforts are not necessarily disseminated by word of mouth simply because users tweet less with an extra followee. Conversely, our results suggest that firms can encourage noncommercial users' tweeting by following them. In addition, firms can benefit more by treating the users differently based on their characteristics. Our findings suggest that following noncommercial users with more active followers is potentially more effective than non-targeting strategies.

6. Concluding Remarks

Social media platforms largely depend on the content generated and shared by their users. Understanding the users' incentives of contribution, more specifically for whom they tweet, lies at the heart of more active online engagement and thus the prosperity of social media platforms. While network effects on Internet-based social media platforms are studied extensively in the literature, we take a step further by showing asymmetric influences of a user's different peer groups. We also complement this strand of literature by systematically studying the asymmetric peer-group size effects on both content generating and sharing.

In this paper we study the asymmetric effects of a user's followee count and follower count on her tweeting intensity on a large-scale social media platform. We propose two methods of approaching the issue of the endogenous network formation. The first is based on a panel method in which we incorporate both user and time fixed effects. The second is a cross-sectional approach in which we use the average observed characteristics of second-order neighbors as instruments. We observe the similar asymmetric effects from both methods. Specifically, a greater population of followers cause users to tweet more while the followee group size has negative effects on tweeting. Further evidence suggests that although asymmetric, the follower effect dominates the followee effect.

Our findings speak to the question: for whom do the users tweet? The estimates indicate that they tweet for followers. Our findings, especially the positive effect of the follower count, indicate that individuals receive benefits or utilities from contributing to online public goods—tweeting in our case. Therefore, our paper provides supporting evidence for the social effects that explain the existence of many public goods with a large number of contributors in online communities. We make managerial recommendations for online platforms to promote more active contributions to online content. In many cases overreaching commercial users or VIPs may lead to reduction in the activeness of users on a platform (the negative "followee" effect). For sustainable development, a social networking platform needs to find a way to get users with far fewer friends to participate. From our simulation results, an effective way would be to promote more followers for these "small" users. In addition, these effects may be moderated by user characteristics, which suggests that the platform should consider user heterogeneity when making decisions. Our results also have implications in terms of marketing strategies on social media platforms.

Our work can be extended in several ways. First of all, future research may enrich our findings by using data that include not only the structure of a user's social network but also the text of tweets and retweets. It would be useful to understand whether the magnitude and direction of our main findings are sensitive to the type of content being generated or shared. Second, beyond the context in this study, future research may apply similar identification strategies we propose to address the endogeneity issue in social networks. Given the existence of large amounts of observational data, future research can explore the structure of social networks and use instruments similar to ours. Finally, we focus on a complete network structure up to a certain point in time. It would enhance our understanding of network effects if future research could explore how the change in network structure, *e.g.* network position, affects a user's contributions to online content. The dynamics of network structure and content generation may also shed light on how they influence each other.

References

Andreoni, J. 2007. Giving Gifts to Groups: How Altruism Depends on the Number of Recipients. *J. Public Econ.* 91(9): 1731-1749.

Aral, S., and D. Walker. 2011. Creating Social Contagion through Viral Product Design: A Randomized Trial of Peer Influence in Networks. *Management Sci.* 57(9): 1623-1639.

Arrow, K. 1972. Gifts and Exchanges. *Philosophy and Public Affairs* 1(4): 343-362.

Becker, G. 1974. A Theory of Social Interactions. *J. Political Econom.* 82(6): 1063-1093.

Bond, R. M., C. J. Fariss, J. J. Jones, A. D. Kramer, C. Marlow, J. E. Settle, and J. H. Fowler. 2012. A 61-Million-Person Experiment in Social Influence and Political Mobilization. *Nature* 489(7415): 295-298.

Bramoullé, Y., Djebbari, H., and B. Fortin. 2009. Identification of Peer Effects Through Social Networks. *J. Econometrics* 150(1): 41-55.

Bramoullé, Y. and R. Kranton. 2007. Public Goods in Networks. *J. Econ. Theory* 135(1): 478-494.

Cameron, A., and P. Trivedi. 2005. *Microeconometrics: Methods and Applications.* Cambridge Univ. Press.

Chamberlin, J. 1974. Provision of Collective Goods as a Function of Group Size. *Amer. Polit. Sci. Rev.* 68(2): 707-716.

Chen, Y., F. M. Harper, J. Konstan, and S. X. Li. (2010). Social Comparisons and Contributions to Online Communities: A Field Experiment on Movielens. *Amer. Econom. Rev.* 100(4): 1358-1398.

Chevalier, J. A., and D. Mayzlin. (2006). The Effect of Word of Mouth on Sales: Online Book Reviews. *J. Marketing Res.* 43(3): 345-354.

Chi, F., and N. Yang. 2011. Twitter Adoption in Congress. *Review of Network Economics* 10(1).

Duan, W., B. Gu, and A. B. Whinston. 2008. Do Online Reviews Matter? An Empirical Investigation of Panel Data. *Decision Support Systems* 45(4): 1007-1016.

Duan, N., W. G. Manning, C. N. Morris, and J. P. Newhouse. 1983. A Comparison of Alternative Models for the Demand for Medical Care. *J. Bus. Econ. Stat.* 1(2): 115-126.

Fehr, E., and A. Falk. 2002. Psychological Foundations of Incentives. *Eur. Econ. Rev.* 46(2002): 687-724.

Gans, J. S., A. Goldfarb, and M. Lederman. 2016. Exit, Tweets, and Loyalty. *Working Paper.*

Ghose, A., and S. P. Han. 2011. An Empirical Analysis of User Content Generation and Usage Behavior on the Mobile Internet. *Management Sci.* 57(9): 1671-1691.

Goes, P., M. Lin, and C. Au Yeung. 2014. "Popularity Effect" in User-Generated Content: Evidence from Online Product Reviews. *Inform. Systems Res.* 25(2): 222-238.

Goldsmith-Pinkham, P., and G. W. Imbens. 2013. Social Networks and the Identification of Peer Effects. *J. Bus. Econ. Stat.* 31(3): 253-264.

Gong, S., J. Zhang, P. Zhao, and X. Jiang 2016. Tweeting Increases Product Demand. *Working Paper.*

Gopinath, S., P. K. Chintagunta, and S. Venkataraman. 2013. Blogs, Advertising, and Local-Market Movie Box Office Performance. *Management Sci.* 59(12): 2635-2654.

Han, B. and L. Yang. 2013. Social Networks, Information Acquisition, and Asset Prices. *Management Sci.* 59(6): 1444-1457.

Huang, Y., P. V. Singh, and A. Ghose. 2013. A Structural Model of Employee Behavioral Dynamics in Enterprise Social Media. *Management Sci.* 61(12): 2825-2844.

Iyer, G., and Z. Katona. 2016. Competing for Attention in Social Communication Markets. *Management Sci.* 62(8): 2304-2320.

Kumar, A., R. Bezawada, R. Rishika, R. Janakiraman, and P. K. Kannan. 2016. From Social to Sale: The Effects of Firm-Generated Content in Social Media on Customer Behavior. *J. Marketing* 80(1): 7-25.

Kumar, V., V. Bhaskaran, R. Mirchandani, and M. Shah. 2013. Creating a Measurable Social Media Marketing Strategy: Increasing the Value and ROI of Intangibles and Tangibles for Hokey Pokey. *Marketing Sci.* 32(2): 194-212.

Lee, Y.-J., K. Hosanagar, and Y. Tan. 2015. Do I Follow My Friends or the Crowd? Information Cascades in Online Movie Ratings. *Management Sci.* 61(9): 2241-2258.

Ma, L., R. Krishnan, and A. L. Montgomery. 2015. Latent Homophily or Social Influence? An Empirical Analysis of Purchase Within a Social Network. *Management Sci.* 61(2): 454-473.

Manjoo, F.. 2016. Breaking up with Twitter. *The New York Times* (November 12), `http://www.nytimes.com/2016/11/13/fashion/breaking-up-with-twitter-presidential-election-2016.html`.

Manski, C. F.. 1993. Identification of Endogenous Social Effects: The Reflection Problem. *Rev. Econ. Stud.* 60(3): 531-542.

McPherson, M., L. Smith-Lovin, and J. M. Cook. 2001. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology* 27: 415-444.

Miller, A. R., and C. Tucker. 2013. Active Social Media Management: The Case of Health Care. *Inform. Systems Res.* 24(1): 52-70.

Murthy, D.. 2012. Towards a Sociological Understanding of Social Media: Theorizing Twitter. *Sociology* 46(6): 1059-1073.

Nam, H., and P. Kannan. 2014. The Informational Value of Social Tagging Networks. *J. Marketing* 78(4): 21-40.

Olson, M. 1965. The Logic of Collective Action. Harvard University Press.

Peng, J., A. Agarwal, K. Hosanagar, and R. Iyengar 2016. Network Embeddedness and Content Sharing on Social Media Platforms. *Working Paper.*

Schweidel, D. A., and W. W. Moe. 2014. Listening In on Social Media: A Joint Model of Sentiment and Venue Format Choice. *J. Marketing Res.* 51(4): 387-402.

Sun, M., and F. Zhu. 2013. Ad Revenue and Content Commercialization: Evidence from Blogs. *Management Sci.* 59(10): 2314-2331.

Toubia, O., and A. T. Stephen. 2013. Intrinsic vs. Image-Related Utility in Social Media: Why do People Contribute Content to Twitter? *Marketing Sci.* 32(3): 368-392.

Wooldridge, J. 2002. Econometric Analysis of Cross Section and Panel Data. The MIT Press.

Wu, L. 2013. Social Network Effects on Productivity and Job Security: Evidence from the Adoption of a Social Networking Tool. *Inform. Systems Res.* 24(1): 30-51.

Xiang, T. 2012. *How Does Douban Make Money?* http://www.technode.com/2012/12/26/how-does-douban-make-money/

Xiao, M. 2010. Is Quality Certification Effective? Evidence from the Childcare Market. *International Journal of Industrial Organization* 28(6): 708-721.

Zhang, J., Y. Liu, and Y. Chen. 2015. Social Learning in Networks of Friends versus Strangers. *Marketing Sci.* 34(4): 573-589.

Zhang, X., and F. Zhu. 2011. Group Size and Incentives to Contribute: A Natural Experiment at Chinese Wikipedia. *Amer. Econom. Rev.* 101(4): 1601-1615.

Zhao, Y., S. Yang, V. Narayan, and Y. Zhao. 2013. Modeling Consumer Learning from Online Product Reviews. *Marketing Sci.* 32(1): 153-169.

Zhu, F., and X. Zhang. 2010. Impact of Online Consumer Reviews on Sales: The Moderating Role of Product and Consumer Characteristics. *J. Marketing* 74(2): 133-148.

Appendix A: Comparing Tencent Weibo with Twitter.com

Compared with Twitter.com, the functions or activities on Tencent Weibo are mostly the same. Users of both platforms can tweet and retweet others' tweets. Both platforms allow users to mention other users by writing their user names following the "@" symbol in tweets or comments. Unlike Tencent Weibo, although Twitter users can also make comments to a tweet by "replying" to it, these replies are private to the two involved parties—the user that posts the tweet and the replier. In sharp contrast, all comments on Tencent Weibo are public (so are tweets) within the platform. Table 11 summarizes the similarities and disctinctions. The Weibo network is identical to that on Twitter, but different from other social media platforms such as Facebook and LinkedIn, where friendships cannot be established without mutual consent and thus is *undirected*.

Table 11 A Comparison of User Activities on Tencent Weibo versus Twitter

User Behaviors		Tencent Weibo	Twitter
Tweets –	Texts of fewer than 140 words, with optional attachment such as pictures and videos.	Yes[a]	Yes
Retweets –	Re-post tweets by others.	Yes	Yes
Tags –	Notifying others of certain tweets (or comments) by typing their user names following the "@" symbol.	Yes	Yes
Comments –	"Commenting" on any tweet.	Public[b]	Private

[a] If the texts are in Chinese, the restriction will be 140 characters instead of words.

[b] Twitter users can also comment on a tweet by "replying" to it. However, unlike the comments on Weibo, the replies on Twitter are private between the communicating parties.

Appendix B: More Summary Statistics

In this appendix we report more summary statistics of our panel and cross-sectional samples. Figure 4 shows the fractions of users with no new tweets between November 30, 2011 and December 11, 2011. It is to note that almost half of the users did not tweet during this period. Figure 5 displays the distributions of new tweet count, followee count, and follower count up to September 15, 2011 after log-transformations.

From the cross-sectional sample, analogous to the longitudinal sample, we also observe skewed distributions for the quantity of tweets and the neighborhood sizes (Table 2 and Figure 6). Therefore, we take log-transformations of these key variables (Figure 6).

Figure 4 **Panel Data – Distribution of Users with respect to the New Total Tweets**

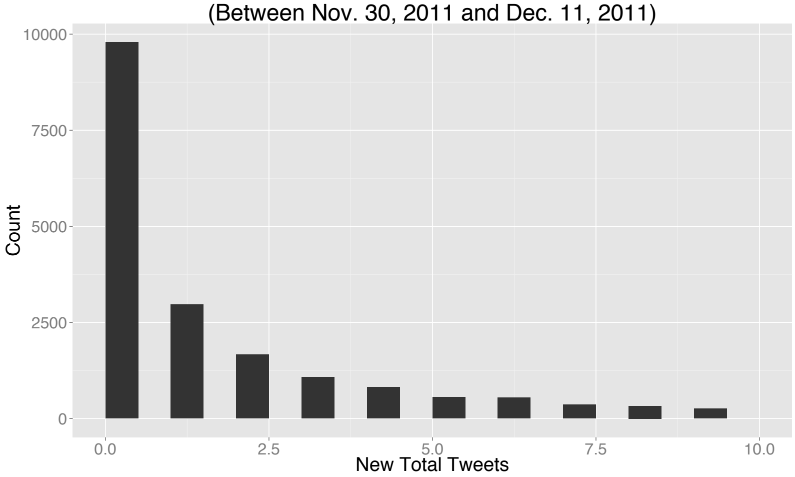


Figure 5 **Panel Data – Distributions of Log-Transformations for Key Variables on September 15, 2011**

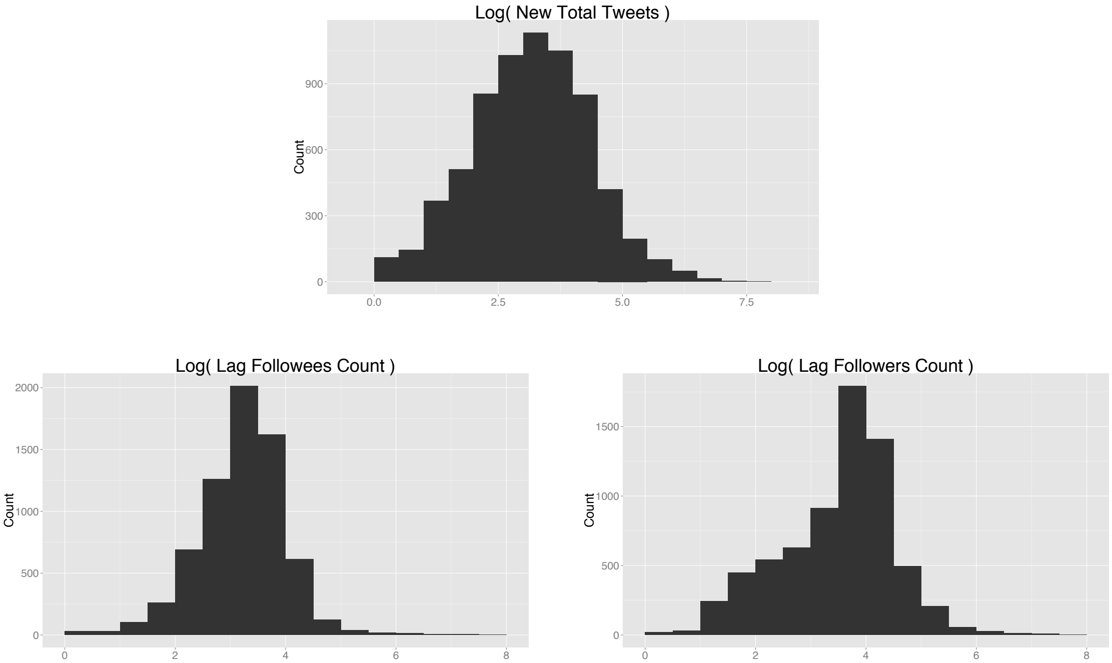
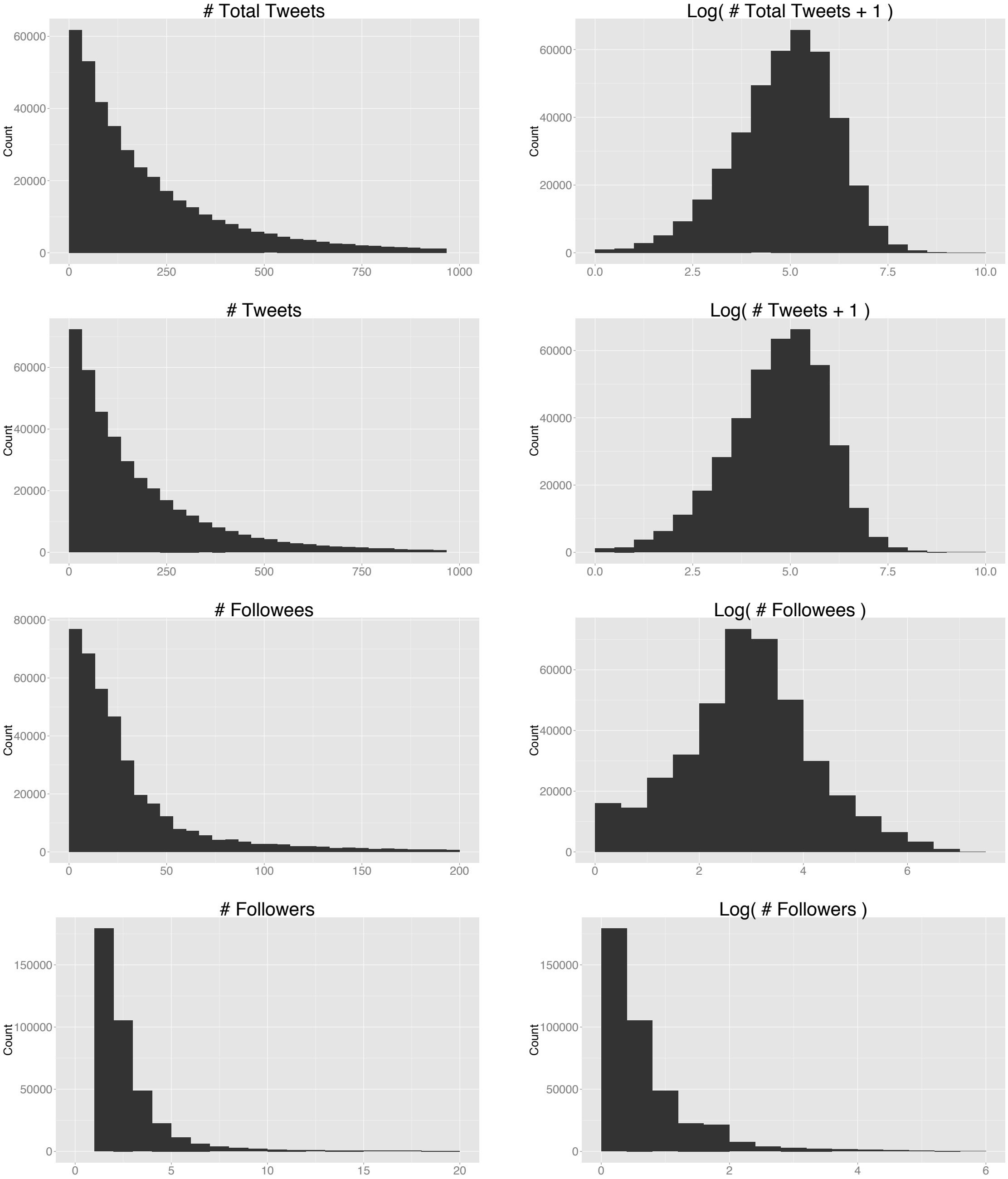


Figure 6 Cross-Sectional Data – Distributions of Key Variables and Log-Transformations



Appendix C: Comparisons of Cross-Sectional Samples

In this appendix, we present an investigation of our main sample in the cross-sectional analyses. Recall that we construct a sample of 402,270 users from the original dataset consisting of 1,392,873 users. We compare these two samples and confirm the representativeness of our main sample. Table below shows summary statistics. We observe that our main sample is largely representative in terms of demographic characteristics, although these users have on average a larger number of followees and of followers and tweet more. This is mainly driven by our IV strategy that focuses on users with at least one second-order followee and one second-order follower.

Table 12 Summary Statistics of the Cross-Sectional Data – Sample Comparisons

	Main Sample			**Original Sample**				
Variables	Med.	Mean	s.d.	Med.	Mean	s.d.	***t*-stats**[a]	***p*-value**
# Total Tweets	143	256.913	447.180	62	142.074	283.663	154.203	< 2.2e-16
# Followees	19	42.427	80.475	11	24.130	53.396	135.864	< 2.2e-16
# Followers	2	17.203	692.496	0	5.162	373.047	10.596	< 2.2e-16
Age	22	24.123	17.147	22	24.571	17.124	-14.589	1
1(Missing Year Birth)	0	0.014	0.117	0	0.016	0.127	-11.004	1
1(Female)	1	0.511	0.500	1	0.504	0.500	7.413	6.173e-14
1(Missing Gender)	0	0.007	0.085	0	0.009	0.096	-12.261	1
1(Age Outliers)	0	0.094	0.291	0	0.098	0.297	-7.977	1
Users		402,470			1,392,873			

[a] We conduct two-sample *t*-test for each variable. The alternative hypothesis is the mean of the corresponding variable in the main sample is *greater* than that in the original sample.

Appendix D: First-Stage Regressions and F-Statistics

In this section, we report the results from the first-stage regressions for our main specification, as shown in the column 4 of Table 6. The OLS estimates, along with F-statistics, of all first-stage regressions are presented in Table 14. The endogenous variables include the neighborhood sizes (followee and follower count) and the average tweets and observed characteristics of all neighbors. We also report the values of F-statistics for all 2SLS regressions in the robustness check (as in Table 8). The results are shown in Table 13. The specifications correspond to those in Table 8.

Table 13 **Values of F-Statistic for 2SLS Regressions in Robustness Checks**

	F-stats for robustness checks[a]			
Specification:	(1)	(2)	(3)	(4)
log (# **Followees**)	2.179e+03	2.179e+03	2.182e+03	2.154e+03
log (# **Followers**)	1.134e+03	1.134e+03	1.132e+03	1.128e+03
Characteristics of All Neighbors				
Avg. Total Tweets			1.836e+03	1.837e+03
Avg. Tweets	1.470e+03			
Avg. Retweets		1.445e+03		
Avg. Age	2.205e+03	2.205e+03	2.229e+03	2.204e+03
Avg. 1(Missing Year Birth)	0.642e+03	0.642e+03	0.641e+03	0.631e+03
Avg. 1(Female)	2.119e+03	2.119e+03	2.128e+03	2.108e+03
Avg. 1(Missing Gender)	0.261e+03	0.261e+03	0.261e+03	0.257e+03
Avg. 1(Age Outlier)	1.201e+03	1.201e+03	1.202e+03	1.209e+03
Num. obs.	402,470	402,470	401,478	400,377

[a] This table presents the F-statistics from the first-stage regressions of our 2SLS regressions in robustness checks. The columns are corresponding to the specifications in Table 8.

Table 14: First-Stage Regressions Results for the 2SLS Estimations of Equation (3) – Main Specification

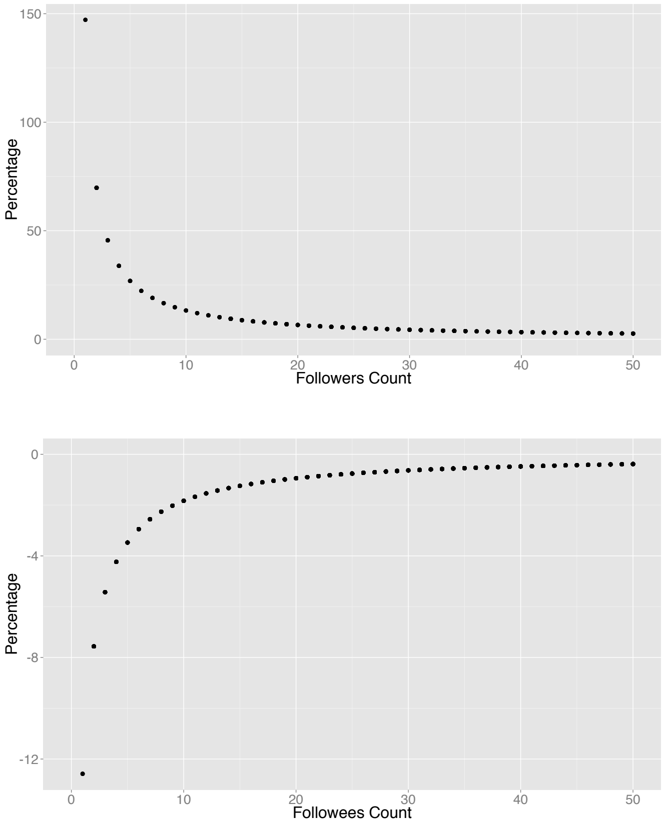
	OLS estimates							
			Average characteristics of all neighbors					
Dep. var.:	log (# Followees)	log (# Followers)	Ttl. Tweets	Age	1(Miss Birth)	1(Female)	1(Miss Gender)	1(Age Outlier)
User Characteristics								
Age	3.545e-04***	0.001***	0.578***	0.044***	-1.645e-04***	-1.299e-04***	-0.406e-04***	0.599e-04***
	(1.228e-04)	(1.059e-04)	(0.044)	(0.00)	(0.685e-07)	(0.174e-04)	(0.643e-07)	(0.111e-04)
1(Missing Year Birth)	-0.216***	-0.156***	25.374***	1.892***	-0.009***	-0.007**	-0.004***	0.006***
	(0.018)	(0.012)	(6.506)	(0.114)	(0.001)	(0.003)	(0.001)	(0.002)
1(Female)	0.042***	-0.037***	34.993***	0.020	0.010***	0.037***	0.002***	-0.007***
	(0.004)	(0.003)	(1.412)	(0.022)	(2.447e-04)	(0.001)	(0.222e-03)	(0.366e-03)
1(Missing Gender)	-0.133***	-0.033	-19.583**	-0.104***	0.003*	0.015***	-0.001	0.001
	(0.026)	(0.023)	(8.214)	(0.149)	(0.002)	(0.004)	(0.001)	(0.003)
1(Age Outliers)	-0.103***	0.085***	20.848***	-0.015	-0.005***	-0.002**	-0.003***	0.007***
	(0.007)	(0.006)	(2.481)	(0.040)	(4.044e-03)	(0.001)	(3.724e-04)	(0.001)
Average Characteristics of Second-Order Followees								
Age	-0.047***	-0.034***	1.916***	0.503***	-0.001***	-0.006***	0.001***	-0.001***
	(0.001)	(0.001)	(0.380)	(0.009)	(0.673e-04)	(2.032e-04)	(0.599e-04)	(1.286e-04)
1(Missing Year Birth)	-2.186***	-1.765***	-1.877e+03***	-20.569***	0.178***	0.389***	0.091***	-0.247***
	(0.138)	(0.056)	(44.140)	(1.076)	(0.010)	(0.024)	(0.008)	(0.017)
1(Female)	-4.103***	-0.927***	-59.135***	-10.141***	-0.003	0.400***	-0.036***	-0.115***
	(0.053)	(0.024)	(14.685)	(0.317)	(0.003)	(0.008)	(0.002)	(0.005)
1(Missing Gender)	-4.634***	-2.983***	-1.482e+03***	14.918***	0.180***	0.452***	0.085***	-0.141***
	(0.164)	(0.069)	(49.624)	(1.242)	(0.012)	(0.027)	(0.010)	(0.018)
1(Age Outliers)	-0.872***	-1.287***	0.245e+03***	-4.632***	-0.087***	-0.079***	-0.008**	0.318***
	(0.100)	(0.042)	(24.753)	(0.575)	(0.004)	(0.013)	(0.004)	(0.009)
Average Characteristics of Second-Order Followers								
Age	0.001***	0.002***	0.710***	0.049***	-0.124e-03***	-1.623e-04***	-0.512***	1.194e-04***
	(1.767e-04)	(0.971e-04)	(0.062)	(0.001)	(0.116e-04)	(0.274e-04)	(0.104e-04)	(0.171e-04)
1(Missing Year Birth)	-0.045	-0.054***	19.407*	1.854***	-0.005***	-0.012***	-0.004**	0.007**
	(0.031)	(0.014)	(10.447)	(0.181)	(0.002)	(0.005)	(0.002)	(0.003)
1(Female)	-0.023***	-0.066***	-11.507***	-0.168***	0.002***	0.009***	0.002***	-0.001
	(0.006)	(0.003)	(2.005)	(0.033)	(3.847e-04)	(0.001)	(3.463e-04)	(0.001)
1(Missing Gender)	-0.048	-0.019	-28.046**	-0.705***	-0.778e-07	0.012*	-0.002	-0.002
	(0.042)	(0.020)	(13.320)	(0.242)	(0.003)	(0.007)	(0.002)	(0.004)
1(Age Outliers)	0.032***	0.058***	14.911***	0.208***	-0.001*	-0.001	-0.001*	0.005***
	(0.011)	(0.005)	(3.649)	(0.065)	(0.001)	(0.002)	(0.001)	(0.001)
(Intercept)	6.499***	2.661***	0.667e+03***	16.348***	0.084***	0.403***	0.042***	0.186***
	(0.064)	(0.034)	(16.636)	(0.431)	(0.003)	(0.009)	(0.003)	(0.006)
F–**stats**	1.617e+03	0.839e+03	1.296e+03	2.174e+03	0.743e+03	2.378e+03	0.190e+03	1.081e+03
Num. obs.	402,470	402,470	402,470	401,478	400,377	402,470	402,470	402,470

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Appendix E: Estimated Marginal Effect of Peer Group Sizes

In this appendix we visualize the estimated marginal effects of the followee count and the follower count based on our main specification (Equation (3)) and coefficient estimates shown in column 4 of Table 6. Specifically, we first calculate the fitted value of the tweet count for each observation in the cross-sectional sample, based on our main specification. Then we calculate the hypothetical values of the predicted tweet count with one extra follower for all observations. The estimated marginal effects are based on the differences between the fitted values. It is straightforward to show that this estimated effect is equivalent to $\left((N_i^r+1)^{\hat{\beta}_2}-(N_i^r)^{\hat{\beta}_2}\right)/(N_i^r)^{\hat{\beta}_2}$, where $\hat{\beta}_2$ is the 2SLS estimate of β_2 as in Equation 3. Note that the effects of a user's characteristics cancel out in the ratio because of the log-log specification.

Figure 7 Estimated Marginal Effects of Neighborhood Sizes on a User's Tweets



The top panel of Figure 7 displays the relationship between the number of followers and the estimated marginal effects (in percentage) on the quantity of tweets with one additional follower. It clearly shows that the estimated effects are positive and decreasing with the number of followers. The bottom panel of Figure 7 presents the relationship between the number of followees and the marginal percentage decrease in the tweet count from one additional followee. Note that the estimated effects are negative for any sizes of followee population. Note also that the two panels in Figure 7 have fairly different scales with the follower effects dominating almost everywhere.

Appendix F: Distributions of New Tweets by the Follower Count

To see that the follower effect is larger at a lower level of the follower count, we dissect how the additional tweets, in the simulation with 1 million new links in Section 5.4.1, are attributed across users with different number of followers by controlling for the number of followees at different levels. More specifically, Table 15 reports the fractions of new tweets generated by users with different number of followers by fixing the number of followees. As an example, column (1) focuses on the subsample of users with fewer than 9 followees, which comprises 25% of the original sample (likewise 19 the median and 40 the 75% quantile). In this subsample, the users with fewer than 1 follower contribute more than 70% of the new tweets, while users in the top 25 percentile (with more than 3 followers) add only about 4% of the new tweets. Similar patterns show up in other subsamples (with different levels of followee counts). We repeat the same process for other simulations in Section 5.4.1 and similar results appear. The follower effects are indeed larger for occasional users who have fewer followers, even controlling for the followee count.

Table 15 Contributions of New Tweets by Users with Different Number of Followers

		Fractional Contributions (%)			
		Controlling for # Followees			
		≤ 9	≤ 19	≤ 40	> 40
# Followers	≤ 1	70.252	65.224	61.480	51.959
	≤ 2	20.133	21.927	22.709	23.314
	≤ 3	5.892	7.504	8.608	10.668
	> 3	3.723	5.345	7.203	14.058