NET Institute*

www.NETinst.org

Working Paper #11-32

October 2011

Ad Revenue and Content Commercialization: Evidence from Blogs

Monic Sun
Stanford University

Feng Zhu
University of Southern California

* The Networks, Electronic Commerce, and Telecommunications ("NET") Institute, http://www.NETinst.org, is a non-profit institution devoted to research on network industries, electronic commerce, telecommunications, the Internet, "virtual networks" comprised of computers that share the same technical standard or operating system, and on network issues in general.

Ad Revenue and Content Commercialization: Evidence from Blogs*

Monic Sun
Stanford University and University of Southern California
Email: monic@stanford.edu

Feng Zhu
University of Southern California
Email: fzhu@marshall.usc.edu

October 1, 2011

*We thank Ron Adner, Victor Bennett, Erik Brynjolfsson, Lisa George, Avi Goldfarb, Yinghua He, Donna Hoffman, Jim Lattin, Albert Ma, Michael Manove, Marc Rysman, Catherine Tucker, Dennis Yao, participants of the *O&S* workshop and the Applied Economics workshop at the University of Southern California, and seminar participants at Stanford University, the University of British Columbia, UC Davis, the 6th Bi-annual Conference of the Software and Internet Industries at the Toulouse School of Economics and the 9th ZEW Conference on the Economics of Information and Communication Technologies for helpful discussion and insightful comments. We are also grateful to Sina.com for providing data, and to *eaglet*, the author of the open source software *Pan Gu Segment*, for answering many questions related to the use of the software. We thank Yinfeng Qin and Rita Ren for research assistance. Finally, we thank the NET Institute (www.NETinst.org) for financial support.

Ad Revenue and Content Commercialization: Evidence from Blogs

Abstract

Many scholars argue that content providers, when incentivized by ad revenue, are more likely to tailor their content to attract "eyeballs," and as a result, popular content may be excessively supplied. We empirically test this prediction by taking advantage of the launch of an ad-revenue-sharing program initiated by a major Chinese portal site in September 2007. Participating bloggers allow the site to run ads on their blogs and receive 50% of the revenue generated by these ads. After analyzing 4.4 million blog posts, we find that, relative to nonparticipants, popular content increases by about 13 percentage points on participants' blogs after the program takes effect. This increase can be partially attributed to topics shifting toward three domains: the stock market, salacious content, and celebrities. We also find evidence that, relative to nonparticipants, participants' content quality increases after the program takes effect.

Key words: ad-sponsored business model, media content, blog, revenue-sharing
(JEL: L82, L86)

1 Introduction

Media consumption today is characterized by three patterns. First, information acquisition is moving online. For example, in 2010, for the first time, more people obtained news online than from print newspapers.[1] Second, a significant portion of the content consumed online is generated by consumers themselves. All the three media sites among the top 10 most-visited websites in the world, YouTube, Blogger and Twitter, are based on user contributions.[2] Third, consumers are increasingly expecting their media consumption to be free.[3] As a result, content providers are under increasing pressure to monetize their content through ad revenue.

Indeed, ad-sponsored business models appear to be increasingly prevalent among online content sites (e.g., Goldfarb 2004). Leading content sites, such as YouTube and Hulu, rely entirely on ad revenue to finance their operations. Small content sites can take advantage of programs from advertising aggregators such as Google's AdSense to make ad revenue without finding advertisers themselves.[4] At the same time, ad-sponsored business models are no longer limited to website owners. Individual content providers today can also earn ad revenue from the content they provide. For example, the most popular video-sharing site, YouTube, started sharing ad revenue with its top contributors in 2007 and has recently extended the ad-revenue-sharing program to all contributors.[5] Many other content sites based on user-generated content (e.g., blog[6] sites such as Blogger and WordPress) have adopted similar practices.

Many scholars criticize the use of ad-sponsored business models in media industries (e.g., Baker 1994; Cross 1994; Herman and McChesney 1997; Turow 1998; Hamilton 2004; McChesney 2004; Anderson and Gabszewicz 2006). They argue that when supported by advertising revenue, media firms, both online and ofline, have incentives to cater content production to

[1] Source: the 2011 State of the News Media report by Pew's Project for Excellence in Journalism (available at http://www.stateofthemedia.org/, accessed March 2011).

[2] Source: www.alexa.com, accessed March 2011.

[3] Wray, Richard. 2010. Media consumption on the increase. Available at http://www.guardian.co.uk/business/2010/apr/19/media-consumption-survey, accessed December 2010.

[4] Website owners can enroll in such programs to enable advertisements on their websites. These advertisements are delivered by advertising aggregators such as Google and generate revenue for the website owners on either a per-click or per-impression basis.

[5] http://www.youtube.com/partners, accessed September 2010. See also Yoganarasimhan (2011) for an interesting discussion on how content propagation depends on the social network structure in a context such as Youtube.

[6] A blog (a blend of the term web log) is a type of website or part of a website. Blogs are usually maintained by an individual with regular entries of commentaries, descriptions of events, or other material, such as pictures or video clips. Entries are commonly displayed in reverse-chronological order.

popular tastes so that they can attract the maximal number of eyeballs. As a result, popular content will be duplicated and excessively supplied, leaving viewers with niche preferences under-served (e.g., Anderson and Gabszewicz 2006). Anecdotal evidence generally supports these criticisms. Broadcast television networks in the US, for example, are frequently blamed for abolishing advertising-unfriendly programs and sticking with redundant ones (e.g., Brown and Cavazos 2003; McChesney 2004; Wilbur 2008). Similarly, most newspapers and magazines are charged with being designed for advertising rather than fundamental editorial content (Bagdikian 2004, pp. 241-246).

To make matters worse, popular content is often not the most consequential to readers and may promote unintended social norms. For example, as violence and sex generally sell well, many content providers routinely employ them, even though the consequences can be socially detrimental (Steinem 1990; Herman and McChesney 1997, p. 137). Although regulations such as the *Fairness Doctrine* require commercial broadcasters to present an ample amount of issues of public importance, these regulations were never enforced (McChesney 2004, p. 44). Because media are important drivers of culture, some critics went so far as to argue that the advertising-media relationship is effectively destroying the culture and that we in society are "amusing ourselves to death" (Postman 2005). The problem is just as severe in developing countries as it is in the developed world (e.g., Zhang 2007). The Chinese speed-dating show, "If You Are the One," for example, was the most popular show in the country in 2010.[7] The show was so salacious, materialistic, and popular that the Chinese government decided to intervene shortly after its launch.[8]

Although scholars repeatedly find support for the claim that ad revenue induces content providers to produce popular or mainstream content (Steiner 1952; Beebe 1977; Spence and Owen 1977; Gal-Or and Dukes 2003; Gabszewicz, Laussel, and Sonnac 2001, 2006; Peitz and Valletti 2008), this prediction has received surprisingly little empirical evaluation. The lack of empirical evidence is perhaps due to the difficulty in establishing a causal relationship: Providers of popular content are more likely to seek ad revenue, and as a result, the causal relationship could be in the opposite direction.[9]

In this study, we empirically evaluate the impact of ad-sponsored business models on con-

[7]CSM Media Research (`http://www.csm.com.cn`) and `http://news.sina.com.cn/m/news/roll/2010-10-16/150221288881.shtml`, both accessed December 2010.

[8]Both the show and the intervention itself gathered tremendous publicity. See an article from the *New York Times* (`http://www.nytimes.com/2010/07/19/world/asia/19chinatv.html`, accessed February 2010) for more details.

[9]In a similar vein, Kind, Nilssen, and Sørgard (2009) show that the degree of content differentiation between media firms' products may affect their dependency on ad revenue.

tent providers' incentives by taking advantage of the introduction of an ad-revenue-sharing program by a major Chinese portal site in September 2007. Participating bloggers allow the site to run ads on their blogs and in return receive 50% of the revenue generated by these ads. We use a difference-in-differences approach to compare content shift of 4,200 participants before and after the program takes effect to that of nonparticipants. We also employ fixed-effects and instrumental-variables approaches to account for bloggers' endogenous decisions to participate in the program. After analyzing 4.4 million blog posts, we find that relative to nonparticipants, popular content increases by about 13 percentage points on participants' blogs after the program takes effect. This increase can be partially attributed to topics shifting toward three domains: the stock market, salacious content, and celebrities. At the same time, we find a significant quality improvement of participants' blog posts.

Our study offers important implications for both consumers and advertising managers. From the consumers' perspective, our results suggest that popular content will be more dominant once ad-revenue-sharing programs take place, which makes it harder for consumers with niche tastes to find free content of their interest. Meanwhile, we also find that that the quality of the posts increases due to the ad-revenue-sharing programs, leading to a potential increase in the readers' welfare. The trade-off hence suggests that the welfare implication of adopting ad-sponsored business model is ambiguous.

For advertisers, it is important to understand how monetary incentives affect the behavior of content providers that they advertise to, which in turn influences the type of consumers that their ads will reach. In addition, as content providers have incentives to generate content that pander to "the lowest common denominator," advertisers need to monitor the type of content their ads are associated with, as such associations may affect their brand images.

It is tempting for intermediaries in media industries such as YouTube and Blogger to adopt ad-revenue-sharing programs. These intermediaries are operating in multi-sided markets, which are prone to winner-take-all dynamics due to significant indirect network effects (e.g., Eisenmann, Parker, and Van Alstyne 2006). As a result, they are under tremendous pressure to grow their traffic by encouraging contributions from content providers and attracting eyeballs. Our results show that indeed sharing ad revenue will incentivize content providers to contribute more frequently. In addition, although consumers may react negatively to ads, sharing ad revenue leads to more popular and high quality content that many consumers enjoy.

On the other hand, intermediaries are likely to see a decrease in the share of content on less popular or niche topics after such ad-revenue sharing programs take place. Consumers with niche tastes may find it harder to find content that matches their interests, and may look for content elsewhere. Furthermore, when multiple intermediaries adopt such ad-revenue-sharing programs, as their content becomes more concentrated on mainstream topics, they will become less differentiated. This loss of differentiation may accelerate the winner-take-all dynamics in such markets. Followers in these markets may be better off without adopting ad-revenue-sharing programs so that they can maintain significant shares of niche content on their sites and stay differentiated.

Broadly speaking, this paper contributes to the growing literature in marketing and economics that examines factors influencing media content. Scholars have examined how the positioning of media content is affected by the entry of national media (e.g., George and Waldfogel 2006) and the mix of consumer types (e.g., George and Waldfogel 2003), how content quality changes with the emergence of the Internet (e.g., Frijters and Velamuri 2010), and how content variety changes as media firms consolidate (e.g., Berry and Waldfogel 2001; George 2002, 2007). They have also identified sources of media bias such as pressure from advertisers or the government (e.g., Price 2003; Reuter and Zitzewitz 2006; Rinallo and Basuroy 2009), and readers' desire for reinforcement of their prior beliefs (e.g., Mullainathan and Shleifer 2005; Gentzkow and Shapiro 2006; Xiang and Sarvary 2007; Gal-Or, Grelani, and Yildirim 2010; Gentzkow and Shapiro 2010). Our paper complements these studies by providing empirical evidence on the impact of ad revenue on the popularity and quality of media content.

The paper proceeds as follows. Section 2 provides details on the empirical setting. Section 3 describes the data. Section 4 presents empirical results and robustness checks, and Section 5 concludes.

2 Background

Our empirical setting is a Chinese portal site, Sina.com, which was founded in 1998. The site today is the 16th most popular website in the world and receives more than 1.4 billion daily page views.[10] It offers many services, including news, emailing, blogging, photo- and video-sharing, microblogging, and instant messaging. Our analysis focuses on its blogging service. Sina started to host blogs for free in September 2005. It is a late mover in the blogging business, as the first Chinese blog-hosting site appeared in 2002,[11] with many other

[10] http://www.alexa.com, accessed February 2011.

[11] http://www.bokee.com, accessed December 2010.

websites providing blogging services since then. For the first two years, the portal site did not place any ads on individual bloggers' content pages. Then, on September 11, 2007, Sina announced an ad-revenue-sharing program. The general public, including bloggers on Sina, were not aware of this program before the announcement, as the company had kept the development of the program strictly confidential to avoid competitive responses by its rivals.

From September 2007 to March 2008, the company conducted a test run of the ad-revenue-sharing program and invited about 3,000 bloggers to participate. About 1,000 bloggers joined the program during this period. In April 2008, the test period ended, and Sina started to accept applications from all bloggers. As indicated in the application guidelines, for an application to be successful, a blog needs to have a minimum of 700 visits per week for four consecutive weeks prior to the application date. Once approved, the site places ads on the blog and the blogger receives 50% of the ad revenue generated by the traffic to her blog pages. To participate in the program, the blogger also needs to provide the site with basic personal information, such as her real name, home address, and bank information. Payments are deposited to participants' bank accounts on a monthly basis whenever the balance exceeds RMB¥100 (equivalent to about US$15).

On the advertiser side, Sina uses a pay-per-impression mechanism: At the beginning of each quarter, it announces a fixed price per thousand impressions, and advertisers decide on the number of impressions to purchase. The site started selling impressions in October 2007, one month after the program's announcement. In November 2007, program participants started to notice ads on their blog pages. Bloggers cannot choose the specific ads to be displayed on their blogs. They receive the same amount of money for each impression at a given time. At the beginning of the program, a blogger would make RMB¥4.5 (equivalent to about US$0.69) per one thousand impressions generated by her blog posts. To avoid annoying viewers, ads are displayed as a small pop-up window in the lower-right corner of the screen, and the pop-up window automatically disappears within 2 to 3 seconds after the web page finishes loading.

The blog-hosting site offers an ideal setting for our study for multiple reasons. First, the site is the largest media website in China. When the ad-revenue-sharing program was introduced, blogs on this site generated about 0.3 billion page views per day, and on a single day, a popular blog post could receive more than 100,000 page views. Given the amount of attention the blogs receive, any systematic change in the content is economically important. Second, unlike many video-sharing sites, our target site offers unlimited storage space to content providers. As a result, bloggers have little incentive to delete their old posts, which allowed us to collect data on the complete history of blog posts from each blogger in our sample. Third, perhaps the most important advantage of our empirical setting is the change in the site's business model: It initially did not compensate content providers, but suddenly introduced the program. The setting hence enables us to observe the change in content production for each participant and estimate the *influence* of ad-sponsored models on the content providers' incentives. Since not every blogger participated in the program, we can use those nonparticipants as a control group in our analysis. Finally, as the site uses a pay-per-impression mechanism on the advertiser side, we do not have to worry about differences among advertisements and the possibility that bloggers tweak their content to target different audiences to get a higher click-through rate.

3 Data

The company provided us with a data set that contains a complete list of all bloggers enrolled in the ad-revenue-sharing program as of January 31, 2009 and the dates each blogger joined the program. Each blogger is associated with a unique ten-digit ID. In total, our data include 5,140 participants, of which 4,200 joined the program after April 2008. We focus our analysis on the bloggers who joined after April 2008, as the motivations of invited participants during the test period could be different. Figure 1 shows the number of bloggers enrolled in the program in each month since April 2008. More than 1,700 bloggers enrolled in the program right after it became open to the general public, and a few hundred bloggers enrolled in the program every month afterwards.

To control for general trends in the content for all blogs, we create a control group by randomly generating another 50 million ten-digit ID numbers. Many of these IDs are mapped to users without blogs: They are users of the portal site's other services. For the bloggers, we first drop those who started blogging after January 2009, and then select those who write more than one blog post per month on average. We apply this last criterion to focus our analysis on active bloggers: Many bloggers only create one or two, often very short, posts right after setting up their blogs and never blog again. It seems that these bloggers want to experience what blogging is like but are not serious about producing any content. In the end, we obtain a list of 26,974 nonparticipants.

We download every blog post that each of the 4,200 participants and 26,974 nonparticipants had written on the site. For each blog post, we collect information on the date it was posted, the title, the number of characters,[12] pictures, and videos in the post, as well as the number of times the post had been read and bookmarked by its viewers. We also collect the tags supplied by the bloggers for each post.

We focus our analysis between May 2007 and January 2009. As we rely on tags to identify popular topics for each month in China, it is critical that we aggregate tags from a sufficient number of posts in each month in our sample. After its launch in September 2005, the site's blogging service experienced accelerated growth in 2006. In the first quarter of 2007, it became the largest blog-hosting site (by the number of visitors) in China. In addition, the site did not introduce the tagging feature until April 2007. In May 2007, around 50% of the blog posts in our sample had tags, and this percentage increased to more than 90% in January 2009. For blog posts with no tags, we use post titles to generate tags.[13]

From May 2007 to January 2009, the bloggers in our dataset composed 4,359,197 blog posts. The eventual participants in the program contributed 1,904,609 (43.7%) posts, and nonparticipants contributed 2,454,588 (56.3%) posts. Figure 2 shows the average number of blog posts in each month by participants and nonparticipants. We find that participants blogged much more frequently on average than nonparticipants. The number of blog posts per month increased for participants over time, and the increase was most pronounced when the program became open to all bloggers. In contrast, the average number of blog posts for nonparticipants declined slightly over time. The pattern suggests that the program motivated participants to produce more content. We also find that the average number of blog posts dropped significantly in February 2008 and January 2009. These drops are most likely due to the Chinese New Year holidays, which typically last 7 to 10 days.[14]

We now consider the popularity of these blog posts. A natural way to consider post popularity is to check whether or not the post is associated with a popular tag. To gauge interest on the tags, we define a tag's popularity in a certain month by the total number of page views of blog posts containing the tag in that month.[15] On average, there are 59,132 tags per month. We rank all the tags based on their popularity and consider the top 150

[12]In Chinese, characters form the basic unit of meaning. Most Chinese words are formed by two or three characters.

[13]We use *Pau Gu Segment*, an open source software that divides Chinese sentences into a set of keywords, to generate these tags. The software is based on a library of more than 170,000 Chinese keywords and has been used by many commercial firms to build Chinese search engines.

[14]The dates for the Chinese New Year in these two years are February 7, 2008 and January 26, 2009. The 7-10 holidays after the Chinese New Year's eve are typically marked by family gatherings and visits to relatives and friends.

[15]It is important to analyze data on a monthly basis, as a tag's popularity may change over time. The tag "Chinese New Year," for example, is popular only at the beginning of a year.

tags in each month as popular tags. We choose this threshold to have a reasonable set of popular tags. Other thresholds, such as 100, 500, or top 0.5% or 1% of all tags, provide similar results. The tag "stock market" is the most popular tag in most months. We then identify blog posts associated with the popular tags in each month as popular posts. On average, 23% of all the posts in our data are classified as popular posts, and these popular posts obtain 63% of the page views.

Next, we compute the percentage of popular blog posts for each blogger in each month. Figure 3 shows how this percentage evolves over time for both participants and nonparticipants in our data. We find that, on average, participants are more likely to post popular content. The percentages of popular content for participants and nonparticipants diverge even more upon the launch of the program: While the percentage of popular content for nonparticipants stayed around 13%, the percentage for participants had a small increase upon the announcement of the program and a significant increase when the program became open to all bloggers. We also notice month-specific effects on the percentages of popular content for both participants and nonparticipants. In May 2008, for example, the percentage of popular content for all bloggers had a sudden increase. This increase resulted from the Wenchuan earthquake, which occurred on May 12, 2008 in China's Sichuan province and killed more than 69,000 people. The earthquake was the most discussed topic in all media in that month, and the tag "earthquake" was the most popular one in that month in our data. The percentage dropped back to its average level for the nonparticipating group right after May 2008 but remained at a high level for the participating group. Similarly, the increase in the percentages of popular content for both groups in August 2008 resulted from the opening of the Summer Olympic Games in Beijing.

4 Regression Analysis

We now turn to regression frameworks to detect shifts in different aspects of the blog content of program participants relative to those of nonparticipants.

4.1 Shift in Content Popularity

We first consider content popularity. We employ a difference-in-differences approach with the specification below:

$$\begin{aligned} \%\ Popular_{it} \quad = \quad & \beta_0 + \beta_1 EventualParticipant_i + \beta_2 EventualParticipant_i \times After_{it} \\ & + \sum_{j=2}^{21} \gamma_j MonthDummy_j + \epsilon_{it}, \end{aligned} \tag{1}$$

where $\%\ Popular_{it}$ is the percentage of popular blog posts among all posts contributed by blogger i in month t. $EventualParticipant_i$ is a dummy that takes the value of 1 if blogger i is an eventual participant in the program, and 0 otherwise. The dummy captures the systematic difference between program participants and nonparticipants. $After_{it}$ is 1 if blogger i is an eventual participant and has already enrolled in the program in month t, and 0 otherwise. β_2 is our difference-in-differences estimator that captures the effect of ad revenue on content popularity for participants. We also include dummies for each month from May 2007 to January 2009 to control for changes in all bloggers' propensity to produce popular content. We cluster the error terms at the blogger level to account for autocorrelation in the data across bloggers and over time (Bertrand, Duflo, and Mullainathan 2004).

We need to address two problems in our specification. First, we need to account for a potential endogeneity problem, as those who apply to join the program are not randomly selected. In other words, some unobserved heterogeneity among bloggers in the error term may be correlated with their decisions to participate in the program, leading to biased estimates. The participants, for example, could in general like to blog about popular topics more than other topics. As a result, their blog posts are more popular, and it is easier for them to qualify for the program.

We take two approaches to address this problem. First, we introduce blogger-level fixed effects to control for time-invariant, unobserved blogger characteristics. Fixed effects allow us to focus on changes in content popularity over time for any given blogger, rather than the absolute levels of content popularity. Fixed effects do not control for time-variant unobservables that may be correlated with the decision to participate in the program. These time-variant unobservables could, for example, lead to different trends over time for participants and nonparticipants. Given this concern, we also construct two instrumental variables by taking advantage of the minimum number of page views required to participate in the program. Valid instruments need to correlate with the decision to participate and affect the dependent variable (*% Popular*) only through the participation decision. Our first instrument is the number of months since a blogger's first post ($BlogAge_{it}$). Our second instrument is the average number of posts per month for the blogger in the past ($AveragePosts_{it}$). The idea is that the longer and the more frequently a blogger has been blogging in the past, the more likely that she would have cultivated an audience base generating more than 700 page views per week, which would make her eligible for the program. At the same time, the two variables are unlikely to be directly correlated with the content of the posts.

Second, our regressions could underestimate the program's impact. Bloggers may tailor their content to improve their chances of getting approved for the program; hence the impact of the program may take effect before they join the program. More generally, as it takes time to increase the popularity of one's blog, some bloggers may choose to shift toward popular content right after the program's announcement and wait until their page views meet the requirement before applying to the program. Indeed, the increase in content popularity for the eventual participants right after the program's announcement, as shown in Figure 3, suggests that such effects may have taken place. Furthermore, some nonbloggers might be incentivized by ad revenue to start blogging and focus on popular content. As a result, bloggers with start dates after September 2007 could be systematically different from those who joined before the program's announcement. Finally, some nonparticipants may also be incentivized by ad revenue. They may have increased their content popularity but still failed to meet the program requirement.

To minimize these effects, we take September 2007 as the breakpoint for all participants and include only those bloggers who started blogging on the site before September 2007. To ensure that the announcement of the program is truly exogenous, we search baidu.com, the top search engine in China, for news related to the ad-revenue-sharing program. All news is dated on or after the day of the program's announcement. We also search the text of all blog posts in our dataset, as bloggers on the site are likely to discuss this program once they become aware of it. All posts that mention this program are also dated after the program's announcement.

Table 1 reports our regression results. In the first three models, we use bloggers' enrollment dates as break points. Model (1) reports the results based on ordinary least square (OLS) regression. On average, a participating blogger's percentage of popular posts before she joins the program is higher than that of a nonparticipant by 22.0 percentage points. This percentage increases by an additional 7.1 percentage points after she joins the program. Model (2) reports the results with fixed effects, which are similar to those in Model

(1). The variable, $EventualParticipant_i$, drops from the regression, as its value does not vary over time. In both Model (1) and (2), the coefficients of the interaction variable, $EventualParticipant_i \times After_{it}$, reflect the average effect of the program on the treated group. Model (3) reports the results with both fixed effects and instrumental variables. The results in Model (3) show that the program's effect becomes stronger after correcting for endogeneity, which suggests that, ceteris paribus, if bloggers were randomly chosen to join the program, the shift toward popular content would have been even greater than what we observe in the actual data. This phenomenon is perhaps because an average blogger's percentage of popular posts is much lower than that of the participants and as a result, there is greater room for her to increase content popularity.

In the next three models, we repeat the analysis in Models (1)-(3) using September 2007 as the break point for all participants. We redefine the dummy variable $After_{it}$ to be 1 if month t is on or after September 2007, and 0 otherwise. We find that the systematic difference in the percentage of popular posts between participants and nonparticipants becomes smaller (15.8%). As expected, the impact of the program is more pronounced: An eventual participant's percentage of popular posts increases by as much as 13.0 percentage points after the program's announcement, which is equivalent to an increase of 65%.[16] The p-value of the over-identification test statistic is 0.26. The results are consistent with our conjecture that many participants started providing popular content in preparation for enrolling in the program after the program's announcement. In the rest of the analysis, we use September 2007 as the break point for all participants.

Table 2 provides first-stage estimation results for Model (6) in Table 1 to illustrate the instrumental variables' relevance. In Models (1) and (2), we include the two instruments, $BlogAge_{it}$ and $AveragePosts_{it}$, separately. In Model (3), we include both of them together. We find that both instruments are highly correlated with becoming a program participant, statistically significant at the 1% level in all three models. The overall Wald Chi-squared test or F-test for the instruments in each model is also highly significant.

4.2 Shift in Content Topics

We now examine the shift in the topics of participants' blog posts after the program takes effect. After speaking with several frequent bloggers in China, we decide to focus on the three most-mentioned topics: the stock market, salacious content, and celebrities. China's stock

[16]Before September 2007, 20% of all posts are classified as popular posts.

market started in early 1990 and has been notorious for its fluctuations.[17] To maximize their returns, many people regularly read blog posts related to the stock market for free opinions and recommendations. Hence, blogging about the stock market is likely to be an effective strategy in attracting traffic. The other two topics, salacious content and celebrities, are universally considered as hot topics.[18]

Two research assistants independently examined the top 150 tags in each month and classified each tag into one of four domains: the stock market, salacious content, celebrities, and others. The results were highly consistent and the few discrepancies were resolved by a meeting of the research assistants. On average, in each month, 12% of the popular tags (e.g., "stock market index" and "stock recommendation") are classified as being related to the stock market, 13% (e.g., "nude photo scandal" and names of Japanese adult-video idols) are classified as being salacious content, 9% (e.g., "celebrity gossip" and names of the celebrities) are classified as being related to celebrities, and 66% (e.g., "earthquake' and 'Chinese New Year') are classified as others.

We identify all blog posts with tags that fall into each of the first three domains. A blog post may be classified into multiple domains. For example, a post on a nude photo scandal involving celebrities is classified as both salacious and related to celebrities. In the end, 6.2%, 6.5%, and 3.0% of all the posts are classified as being related to the stock market, salacious content, and celebrities, respectively. We then compute the percentage of posts in each of these three domains for each blogger in each month, and use these three percentages as dependent variables to repeat the difference-in-differences analysis. Table 3 reports the regression results.

For each of the three percentages above, we report the results with fixed effects, and the results with both fixed effects and instrumental variables. The results demonstrate significant shifts of content toward all three domains. In total, the blog posts in these three domains for participants increase by 6.6% percentage points (based on the specification with both fixed effects and instrumental variables) relative to nonparticipants.

4.3 Shift in Content Quality

Finally, we consider the impact of ad revenue on content quality. As high-quality content attracts eyeballs, participating bloggers may devote more effort to improving the quality

[17]See, for example, http://www.businessweek.com/globalbiz/content/jan2010/gb2010016_835230.htm, accessed March 2011.

[18]Although it is illegal in China to post images or videos that contain nudity or text containing explicit descriptions of sexual acts, bloggers can include images or text that are sexually suggestive.

of their posts. To identify the extent of such effects, we develop several measures on post quality. For each post, we first compute the percentage of viewers who bookmark the post as one of their favorites.[19] Bloggers cannot tell who bookmarked their posts and hence cannot reciprocate by visiting the blogs of their patrons. Therefore, the only benefit of bookmarking a post is the convenience of re-accessing it in the future. For each blogger in each month, we compute the average of the percentages for all her posts and denote this measure as $\%\ Bookmark_{it}$, which reflects the utility that readers receive from reading the post.

We also measure the degree of each blogger's effort by the average number of characters, pictures, and video clips in her posts. For any given blogger, the more effort she devotes to writing, the more likely it is that the blog post has a higher quality. In general, pictures and videos make a post more attractive, although they may also require more effort from the blogger. We denote these measures as $Num\ Chars_{it}$, $Num\ Pics_{it}$, and $Num\ Videos_{it}$, respectively. We take the logarithm of the average number of characters to minimize the effects of outliers.

We apply the same difference-in-differences approach using each of the four measures mentioned above as the dependent variable. Table 4 reports the results. We find significant improvement for all measures after the program's announcement. The significant increase in $\%\ Bookmark_{it}$ is noteworthy: one may have thought that blog posts would have become more repetitive and less worthy of a bookmark, Instead, we find that the average percentage of readers who bookmark a post, $\%\ Bookmark_{it}$, increases rather than decreases.

4.4 Exploring Robustness[20]

The blog-hosting site's sudden announcement of the program gives us an opportunity to employ a useful falsification test. If our assumption of orthogonality between blogger-specific unobservables and their decisions to participate in the program is violated after employing blogger-level fixed effects and instrumental variables, our data will produce diverging patterns for participants and nonparticipants in periods even before the program's announcement. We thus regress each outcome variable in Tables 1, 3 and 4 on dummies for each month between May 2007 to August 2007 and their interactions with $EventualParticipant_i$. We find little evidence that the trends for participants and nonparticipants differ significantly prior to the program's announcement in fixed-effects specifications with or without instrumental

[19] For each blog post, the site provides a button that any reader with an account on the site can use to bookmark the blog post, which puts the post in her personal collection.

[20] Results for robustness checks are not reported.

variables. The absence of such false positives increases our confidence that the observed shifts for participants are caused by the program.

We are also concerned about the potential strategic manipulation of tags by program participants. For example, participants may use more tags for their posts to attract more readers after the program's announcement. They may also supply popular tags even though these tags are not accurate descriptions of their posts. Such strategic manipulation could have contributed to our finding that program participants are more likely to produce popular content or content related to one of the three topic domains. To address this concern, we generate tags for all posts based on the text in each post.[21] For each blog post, we tokenize the text into individual words and use the four most frequently mentioned nouns as the tags.[22] We follow the same procedure to identify popular posts and the domains to which each popular post belongs, and then repeat the analysis in Tables 1 and 3. We obtain similar results.

Finally, we are concerned that some bloggers may write much more frequently than others, and as a result, their blog posts may have a disproportionate influence in determining whether certain content is popular or not: If a small number of prolific bloggers mention the same tag in every blog post they write, this tag is likely to be classified as a popular tag even if none of the other bloggers use the tag. We repeat our analysis after excluding bloggers whose average monthly number of posts is more than four standard deviations above the mean. In total, 133 bloggers are dropped from the analysis. We obtain similar results.

5 Discussion and Conclusion

In this paper, we attempt to empirically evaluate the impact of ad-sponsored business models on the incentives of content providers online. We find that, consistent with the theoretical literature, content providers sponsored by ad revenue are more likely to generate content that pander to the lowest common denominator. Meanwhile, we also find that ad-sponsored models lead to increased effort toward generating content and making it more likeable. The welfare implication of adopting an ad-sponsored business model is therefore ambiguous.

Our study has several limitations. First, in our empirical setting, content providers do not face a capacity constraint, but in offline media, such as television and newspapers, the capacity constraint can be critical. In such cases, the substitution of unpopular content with

[21]One disadvantage of this approach is that we may not analyze blog posts correctly if they contain mostly pictures or videos.

[22]If a post contains fewer than four nouns, we use all the nouns as tags.

popular content may be more pronounced when content providers rely more on ad revenue.

Second, in China, pornography and politically controversial issues are generally not allowed. Therefore, some of our results could be different in other cultural settings.

Third, content providers in our setting are compensated by the number of ad impressions they serve. It will be interesting to compare our results to settings where content providers are paid by the number of clicks on ads (e.g., Ghose and Yang 2009; Goldfarb and Tucker 2011). In such cases, while content providers will still have incentives to maximize the number of eyeballs, they may also try to match their content with advertisements to increase readers' incentives to click on ads.[23]

In addition, several studies have shown that user-generated content such as blogs may influence consumers' behavior (e.g., Dewan and Ramaprasad 2010; Gopinath, Chintagunta, and Venkataraman Forthcoming). As ad revenue leads to shifts in content topics and quality, future research can investigate how such shifts affect content readers. For example, as more bloggers are discussing stock markets, are readers receiving better stock recommendations, and as a result, making wiser investment decisions?[24] Similarly, are celebrities becoming more popular among these readers? Another interesting question along this dimension has to do with consumers' perception of the websites dominated by popular contents and ads. They can either be discouraged by the fact that the sites are created for ad revenue and hence become less motivated in reading the content (Porter and Donthu 2008), or be encouraged by the fact that other readers are also interested in the content (Tucker and Zhang 2010).

Finally, our work provides implications for future theoretical studies of ad-sponsored business models. In prior studies examining location choices under ad-sponsored business models, quality is often assumed to be exogenous and identical for competing content providers (e.g., Gabszewicz, Laussel, and Sonnac 2006; Godes, Ofek, and Sarvary 2009). As a result, these models do not predict the effects of ad-sponsored business models on content quality. Our results suggest that both content quality and location choices need to be endogenized to fully understand the impact of ad-sponsored business models.

[23] See Zettelmeyer (2000) for a detailed discussion on firms' pricing and communications strategies when they compete in both online and offline channels.

[24] To explore the consequences of content shifts, one may need to consider content providers' incentives to directly copy and paste others' popular content to attract traffic. See Desai, Purohit, and Vernik (2010) for a discussion on the effect of digital right management in the context of digital music.

References

Anderson, Simon P., Jean J. Gabszewicz. 2006. The media and advertising: A tale of two-sided markets. Victor A. Ginsburgh, David Throsby, eds., *Handbook of the Economics of Art and Culture*, vol. 1, chap. 18. Elsevier B.V., 568–614.

Bagdikian, Ben H. 2004. *The New Media Monopoly*. Beacon Press.

Baker, C. Edwin. 1994. *Advertising and a Democratic Press*. Princeton, NJ: Princeton University Press.

Beebe, Jack H. 1977. Institutional structure and program choices in television markets. *Quarterly Journal of Economics* **91**(1) 15–37.

Berry, Steven T., Joel Waldfogel. 2001. Do mergers increase product variety? Evidence from radio broadcasting. *Quarterly Journal of Economics* **116**(3) 1009–1025.

Bertrand, Marianne, Esther Duflo, Sendhil Mullainathan. 2004. How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics* **119**(1) 249–275.

Brown, Keith S., Roberto J. Cavazos. 2003. Empirical aspects of advertiser preferences and program content of network television. Federal Communications Commission Media Bureau Staff Research Paper 2003-1.

Cross, Donna W. 1994. *Media Speak: How Television Makes Up Your Mind*. Penguin Group.

Desai, Preyas, Debu Purohit, Dinah Vernik. 2010. Digital music set free: The flip side of DRM protection. Working paper, Duke University.

Dewan, Sanjeev, Jui Ramaprasad. 2010. Consumer blogging and music sampling: Long tail effects. Working paper, UC Irvine Paul Merage School of Business.

Eisenmann, Thomas R., Geoffrey Parker, Marshall W. Van Alstyne. 2006. Strategies for two-sided markets. *Harvard Business Review* **84**(10) 92–101.

Frijters, Paul, Malathi Velamuri. 2010. Is the Internet bad news? The online news era and the market for high-quality news. *Review of Network Economics* **9**(2).

Gabszewicz, Jean J., Dider Laussel, Nathalie Sonnac. 2001. Press advertising and the ascent of the 'pensée unique'. *European Economic Review* **45**(4-6) 641–651.

Gabszewicz, Jean J., Didier Laussel, Nathalie Sonnac. 2006. Competition in the media and advertising markets. *Manchester School* **74**(1) 1–22.

Gal-Or, Esther, Anthony J. Dukes. 2003. Minimum differentiation in commercial media markets. *Journal of Economics & Management Strategy* **12**(3) 291–325.

Gal-Or, Esther, Tansev Grelani, Tuba Pinar Yildirim. 2010. The impact of advertising on media bias. Working paper.

Gentzkow, Matthew, Jesse M. Shapiro. 2006. Media bias and reputation. *Journal of Political Economy* **114**(2) 280–316.

Gentzkow, Matthew, Jesse M. Shapiro. 2010. What drives media slant? Evidence from U.S. daily newspapers. *Econometrica* **78**(1) 35–71.

George, Lisa M. 2002. Ownership concentration and product variety in daily newspaper markets. Lorrie F. Cranor, Shane Greenstein, eds., *Communications Policy and Information Technology: Promises, Problems, Prospects*, chap. 9. MIT Press, Cambridge, MA, 235–253.

George, Lisa M. 2007. What's fit to print: The effect of ownership concentration on product variety in daily newspaper markets. *Information Economics and Policy* **19**(3-4) 285–303.

George, Lisa M., Joel Waldfogel. 2003. Who affects whom in daily newspaper markets? *Journal of Political Economy* **111**(4) 765–785.

George, Lisa M., Joel Waldfogel. 2006. The "New York Time" and the market for local newspapers. *American Economic Review* **96**(1) 435–447.

Ghose, Anindya, Sha Yang. 2009. An empirical analysis of search engine advertising: Sponsored search in electronic markets. *Management Science* **55**(10) 1605–1622.

Godes, David, Elie Ofek, Miklos Sarvary. 2009. Content vs. advertising: The impact of competition on media firm strategy. *Marketing Science* **28**(1) 20–35.

Goldfarb, Avi. 2004. Concentration in advertising-supported online markets: An empirical approach. *Economics of Innovation and New Technology* **13**(6) 581–594.

Goldfarb, Avi, Catherine Tucker. 2011. Search engine advertising: Channel substitution when pricing ads to context. *Management Science* **57**(3) 458–470.

Gopinath, Shyam, Pradeep K. Chintagunta, Sriram Venkataraman. Forthcoming. Measuring local market-level differences in the effects of pre-release blogs on a movie's opening box-office performance. *Marketing Science* .

Hamilton, James T. 2004. *All the News That's Fit to Sell*. Princeton, NJ: Princeton University Press.

Herman, Ed S., Robert W. McChesney. 1997. *The Global Media: The Missionaries of Global Capitalism*. Cassell.

Kind, Hans Jarle, Tore Nilssen, Lars Sørgard. 2009. Business models for media firms: Does competition matter for how they raise revenue? *Marketing Science* **28**(6) 1112–1128.

McChesney, Robert W. 2004. *The Problem of the Media: U.S. Communication Politics in the Twenty-First Century*. Monthly Review Press.

Mullainathan, Sendhil, Andrei Shleifer. 2005. The market for news. *American Economic Review* **95**(4) 1031–1053.

Peitz, Martin, Tommaso M. Valletti. 2008. Content and advertising in the media: Pay-tv versus free-to-air. *International Journal of Industrial Organization* **26**(4) 949–965.

Porter, Constance Elise, Naveen Donthu. 2008. Cultivating trust and harvesting value in virtual communities. *Management Science* **54**(1) 113–28.

Postman, Neil. 2005. *Amusing Ourselves to Death: Public Discourse in the Age of Show Business*. Penguin.

Price, Cindy J. 2003. Interfering owners or meddling advertisers: How network television news correspondents feel about ownership and advertiser influence on news stories. *Journal of Media Economics* **16**(3) 175–188.

Reuter, Jonathan, Eric Zitzewitz. 2006. Do ads influence editors? Advertising and bias in the financial media. *Quarterly Journal of Economics* **121**(1) 197–227.

Rinallo, Diego, Suman Basuroy. 2009. Does advertising spending influence media coverage of the advertiser? *Journal of Marketing* **73**(6) 33–46.

Spence, Michael, Bruce Owen. 1977. Television programming, monopolistic competition, and welfare. *Quarterly Journal of Economics* **91**(1) 103–126.

Steinem, Gloria. 1990. Sex, lies and advertising. *Ms.* **July/August** 18–28.

Steiner, Peter O. 1952. Program patterns and preferences, and the workability of competition in radio broadcasting. *Quarterly Journal of Economics* **66**(2) 194–223.

Tucker, Catherine, Juanjuan Zhang. 2010. Growing two-sided networks by advertising the user base: A field experiment. *Marketing Science* **29**(5) 805–814.

Turow, Joseph. 1998. *Breaking Up America: Advertisers and the New Media World*. University Of Chicago Press.

Wilbur, Kenneth C. 2008. A two-sided, empirical model of television advertising and viewing markets. *Marketing Science* **27**(3) 356–378.

Xiang, Yi, Miklos Sarvary. 2007. News consumption and media bias. *Marketing Science* **26**(5) 611–628.

Yoganarasimhan, Hema. 2011. Impact of social network structure on content propagation - A study using YouTube data. Working paper, UC Davis.

Zettelmeyer, Florian. 2000. Expanding to the internet: Pricing and communications strategies when firms compete on multiple channels. *Journal of Marketing Research* **37**(3) 292–308.

Zhang, Bo. 2007. On the vulgarization of mass media. *Press Circles* **4** 66–68.

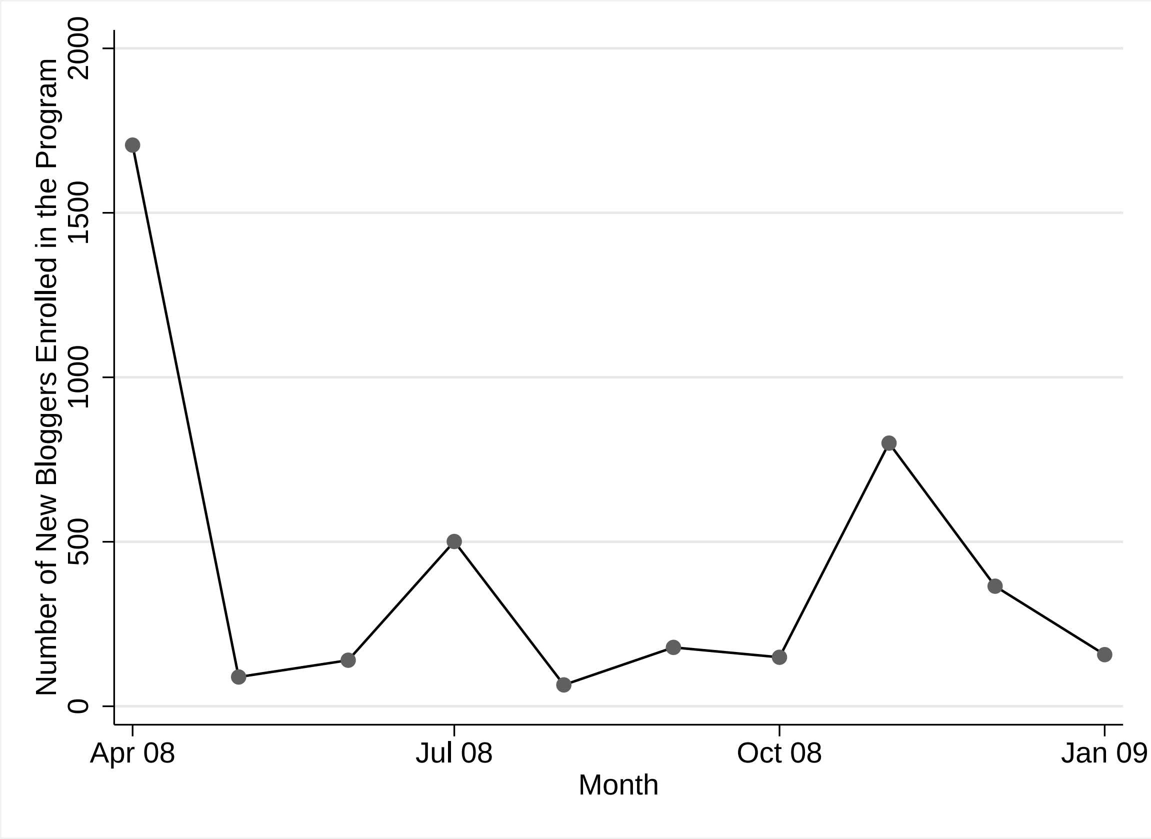
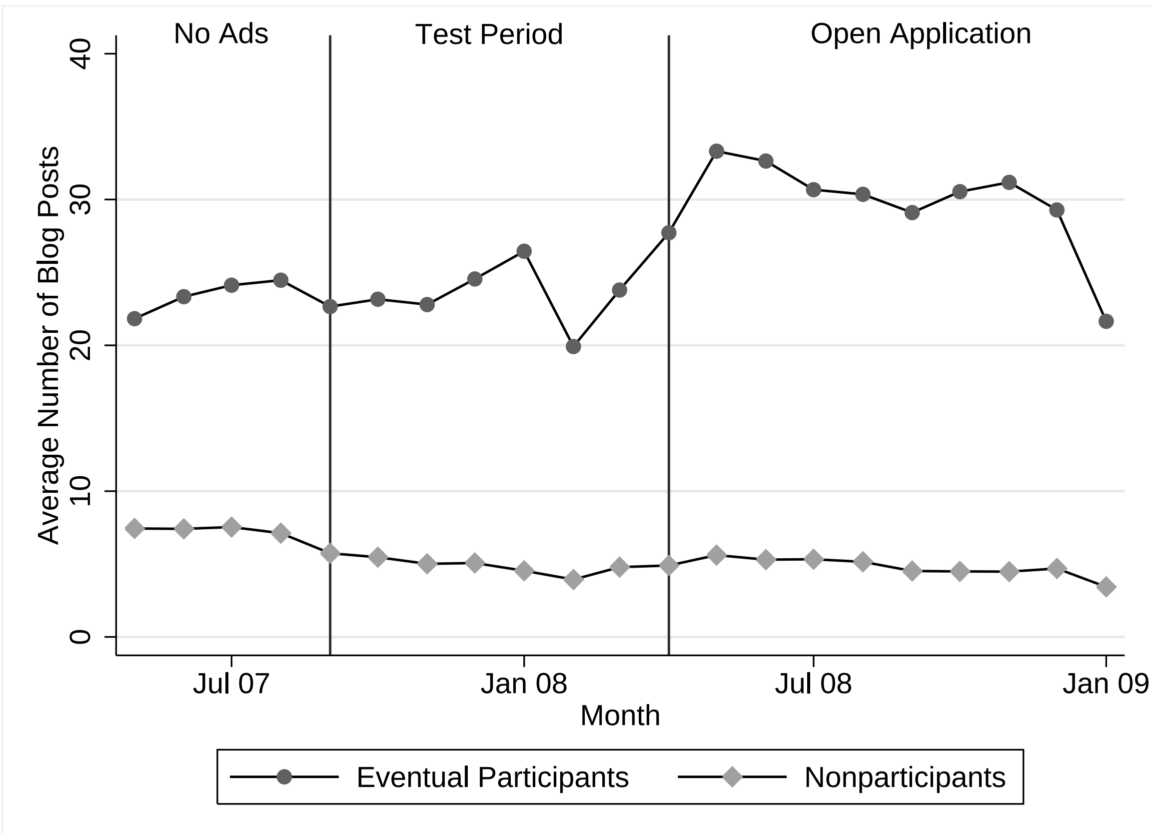
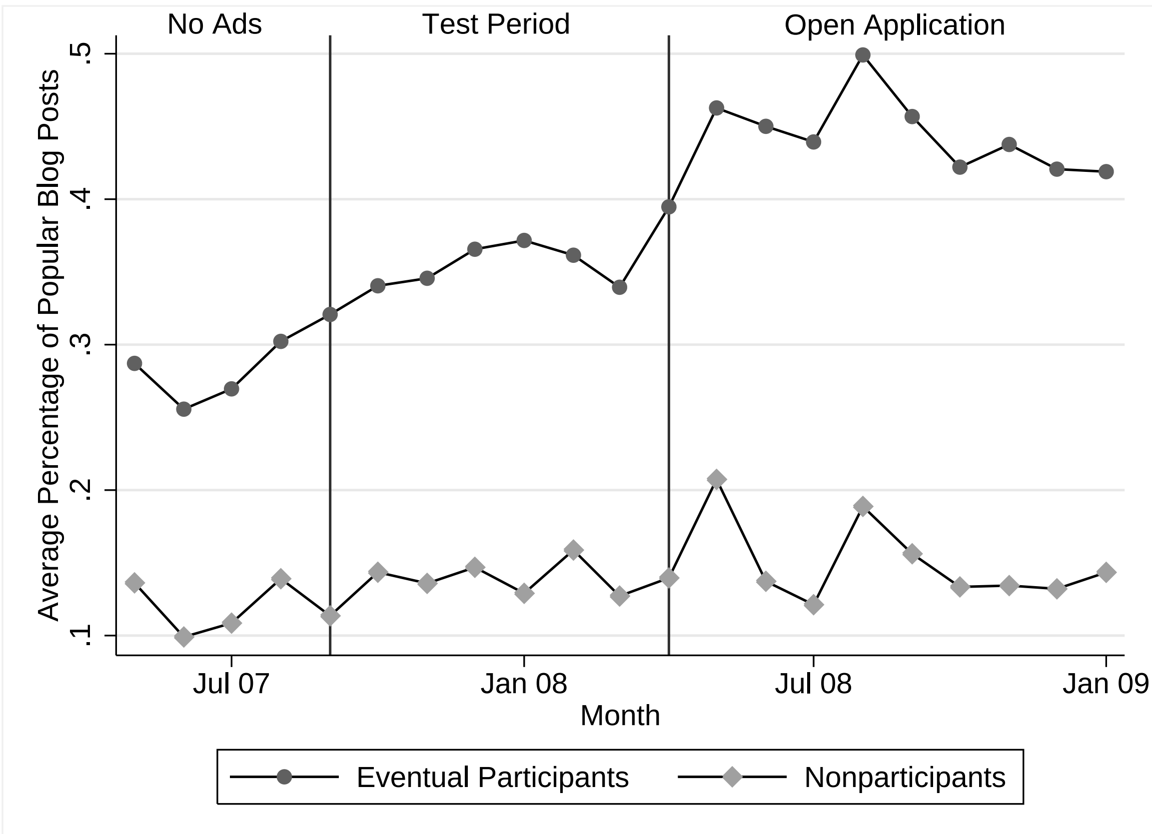


Figure 1: Number of Bloggers Enrolled in the Program in Each Month



Note: The two vertical lines indicate September 2007 and April 2008, respectively.

Figure 2: Average Number of Blog Posts in Each Month



Note: The two vertical lines indicate September 2007 and April 2008, respectively.

Figure 3: Average Percentage of Popular Blog Posts in Each Month

Table 1: The Impact of Revenue Sharing on Content Popularity

	Enrollment dates as Break Points			9/2007 as the Break Point		
Model	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable	% Popular	% Popular	% Popular	% Popular	% Popular	% Popular
EventualParticipant	0.220***			0.158***		
	(0.004)			(0.005)		
EventualParticipant	0.071***	0.070***	0.078***	0.096***	0.093***	0.130***
× After	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.017)
Month Dummies	Yes	Yes	Yes	Yes	Yes	Yes
Observations	544,209	544,209	544,209	448,973	448,973	448,973
Adjusted R-squared	0.097	0.016	0.016	0.093	0.017	0.016
Number of IDs		31,174	31,174		21,792	21,792
Specification	OLS	FE	FE/2SLS	OLS	FE	FE/2SLS

Note: Heteroskedasticity-adjusted standard errors in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 2: First Stage Regressions and Instrument Relevance

Model	(1)	(2)	(3)
Dependent Variable	EventualParticipant × After	EventualParticipant × After	EventualParticipant × After
BlogAge × After	0.080***		0.052***
	(0.003)		(0.005)
AveragePosts × After		0.009***	0.008***
		(0.001)	(0.001)
Month Dummies	Yes	Yes	Yes
Observations	448,973	448,973	448,973
Adjusted R-squared	0.144	0.204	0.209
Number of IDs	21,792	21,792	21,792
F-statistics	154.4	173.1	167.1
Specification	FE	FE	FE

Note: Heteroskedasticity-adjusted standard errors in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 3: The Impact of Revenue Sharing on Content Topics

Model	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable	Stock Market	Stock Market	Salacious Content	Salacious Content	Celebrities	Celebrities
EventualParticipant	0.012***	0.024**	0.024***	0.026***	0.014***	0.016*
× After	(0.002)	(0.012)	(0.002)	(0.007)	(0.002)	(0.009)
Month Dummies	Yes	Yes	Yes	Yes	Yes	Yes
Observations	448,973	448,973	448,973	448,973	448,973	448,973
Adjusted R-squared	0.002	0.001	0.003	0.003	0.001	0.001
Number of IDs	21,792	21,792	21,792	21,792	21,792	21,792
Specification	FE	FE/2SLS	FE	FE/2SLS	FE	FE/2SLS

Note: Heteroskedasticity-adjusted standard errors in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 4: The Impact of Revenue Sharing on Content Quality

Model	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent Variable	% Bookmarks	% Bookmarks	Num Chars	Num Chars	Num Pics	Num Pics	Num Videos	Num Videos
EventualParticipant	0.055***	0.096***	1.221***	1.684***	1.490***	2.096***	0.001***	0.001*
× After	(0.003)	(0.010)	(0.036)	(0.150)	(0.040)	(0.138)	(0.000)	(0.001)
Month Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	448,973	448,973	448,973	448,973	448,973	448,973	448,973	448,973
Adjusted R-squared	0.006	0.004	0.050	0.049	0.015	0.013	0.003	0.003
Number of IDs	21,792	21,792	21,792	21,792	21,792	21,792	21,792	21,792
Specification	FE	FE/2SLS	FE	FE/2SLS	FE	FE/2SLS	FE	FE/2SLS

Note: In Models (1) and (2), we multiple the dependent variable by 100 for the ease of displaying coefficients. We take the logarithm of *Num Chars* to minimize the impact of outliers. Heteroskedasticity-adjusted standard errors in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.