Review of the paper “Ad Revenue and Content Commercialization: Evidence from Blogs” by Sun and Zhu (2013)

Course: E-Business and Online Marketing

Instructor: Prof. Stefan Lessmann

Students: Tyler Marangi, Christian Geike

Handed -in: 21.08.2015

The paper "Ad Revenue and Content Commercialization: Evidence from Blogs" by Sun and Zhu (2013) investigates the relationship between pay-per-impression ad schemes with revenue sharing and blogger behavior in China. If, as the paper purports to demonstrate, such schemes do indeed influence blogger behavior directly, then they will need to be considered a dynamic factor to be accounted for in marketing management. The data-driven nature of the approach makes the paper well-suited for the Management Science journal. Management Science covers many topics, including marketing, with an emphasis on “the foundational disciplines of economics, mathematics, psychology, sociology, and statistics” (http://pubsonline.informs.org/journal/mnsc#), making this a particularly eligible article that cuts across many of the dimensions of interest for the journal.

Our paper review is structured as follows: First, we provide a detailed overview of the paper, focusing on its research design and its results. Second, we comment on the specific contribution of the paper and its key implications for academia and industry. Third, we give our overall assessment of the paper. Fourth, we elaborate on our overall assessment by providing detailed comments to each of the points we have identified. Finally, some minor comments complete the review.

A) Detailed overview of the paper

The authors analyze data from the blogging service of Sina.com, being the 16th most popular website in the world with 1.4 billion daily page views. Founded in 1998, its blogging service started in 2005. Having introduced the feature that bloggers can provide tags to their blog posts in May 2007, Sina.com soon announced an ad revenue sharing program on 11 September 2007 whereby participating bloggers receive 50% of the ad revenue generated by the traffic on their blogs. Advertisers, on the other side, have to pay for their ads, which are displayed in a pop-up window in the lower right corner of the screen for about two to three seconds, on a pay-per-impression mechanism: at the beginning of each quarter, they can decide about the number of impressions to purchase whereby the price for 1000 impressions is fixed by the seller Sina.com. There is no price discrimination between different advertisers. Also, the ads are randomly allocated to the participating blogs so that these bloggers have no incentive to twist their blog posts to specific ads in an attempt to increase own page views.

After a test period from 11 September 2007 till end of April 2008 during which 940 bloggers started to participate on the basis of an invitation by Sina.com, the open application period began in May 2008. The criteria to be eligible for becoming a participating blogger in the program is to have at least 700 average weekly page views for the last four consecutive weeks. Since the authors were provided with a list of all bloggers enrolled in the program and their enrollment date by Sina.com as of 31 January 2009, their research design comprises three time periods: a pre-announcement period beginning in May 2007, a post-announcement period from 11 September 2007 till April 2008 and an open-application period from May 2008 till January 2009.

Out of the 5240 participating bloggers, the authors take those 4200 bloggers into their analyses who enrolled during the open application period, i.e. they excluded those starting in the test period because, being invited to join the program, they could have a different motivation. As a control group, they take all the nonparticipants writing more than one post per month on average. The idea behind that criterion is to have serious bloggers. They end up with 26,974 nonparticipants. Altogether, the authors saved and analyzed all the 4,359,197 blog posts of these participating and nonparticipating bloggers during the three time periods: 1,904,609 (43.7%) for participants and 2,454,588 (56.3%) for nonparticipants. The data saved for each blog post are: date of blog post; its title; number of characters, pictures and videos; number of times the post has been read; number of bookmarks and the tags supplied by the bloggers. Bloggers could supply their blog posts with as little or many tags as they like, having no upper or lower tag limit to meet. Since the introduction of the tag feature in May 2007, the rate of blog posts provided with tags increased from 50% to 90% till January 2009. For all the untagged blog posts, the authors generate tags from the post titles using "Pan Gu Segment", an open source software that has also been used commercially to build Chinese search engines. As seen in figure 2 in the paper, eventual participants blogged much more frequently on average, a trend that even increases over time as the program gets announced and started, while for nonparticipants, a slight decline can be observed on average.

On the basis of these collected and – with respect to missing tags- generated data, the authors apply model-free descriptive analyses as well as different regression analyses in order to investigate the following topics: 1) shift in content popularity, 2) shift in content topics, 3) shift in content quality and 4) program effects across participants and over time.

<u>1) Shift in content popularity</u>

To define tag popularity, the authors first found out how often each tag is contained within the blog posts for each month. Ranking all the tags, they define the first 150 tags, being incorporated most often within blog posts, as popular for that month. The monthly focus results from the possibility that what is considered as popular can change over time. Based on the 150 popular tags for each month, all the blogs post for that month being associated with at least one of these popular tags are identified as popular posts. Taking all the months together, 23% of all blog posts are classified as popular, generating 63% of all page views. Next, the authors calculate the percentage of popular posts of each blogger in each month. Aggregating the results over all the bloggers and over time, they find that on average, 16.9% of bloggers post popular content. Subdividing between eventual participants and nonparticipants, as seen in figure 4 and table 2 in the paper, one sees that eventual participants have a mean percentage of popular posts of 27.9%, 33.1% and 41.2% before the announcement, in the test period and in the open application period, respectively, while nonparticipants' mean percentage of popular posts stays at 12.1%, 12.4% and 12.5% within the respective periods. Apart from the higher percentage rate for eventual participants, the gap between the two means widens over time, suggesting that eventual participants increase their share of popular content after program announcement and launch.

Next, the authors turn to different regression models based on equation 1 in the paper. While the dependent variable is the percentage of popular posts for each blogger in each month, the estimated coefficients ß1 and ß2 capture the systematic difference between participants and nonparticipants and, most importantly, the program effect of ad revenue sharing on content popularity for participants, respectively. The authors include monthly dummies to control for changes of all bloggers' propensity for posting popular content, as well as blogger-level error terms, accounting for autocorrelation across bloggers and over time. This first basic OLS regression model is extended to account for the potential endogeneity problem that program participants self-select themselves, i.e. that their decision to participate in the program might be systematically influenced by some unobserved characteristics. Blogger-level fixed effects are introduced within a second regression model to control for those unobserved characteristics that are time-invariant, while two additional instrumental variables, Blogging_Age(it), i.e. the number of month since a blogger's first post, and Blogging_Freq(it), i.e. the average number of posts per month for the blogger in the past, are introduced in addition to the fixed effects within a third two-stage least-square regression model to control for those unobserved characteristics that are time-variant. Apart from this self-selection problem, the authors identify the problem that bloggers might change their content before enrolling for the program in an attempt to make their blog more popular and meet the program requirement; this can happen from as early as program announcement. Considering eventual participants only after the point of enrollment, the first three regression models could consequently underestimate the impact of the program. Also, new bloggers starting only after the program announcement on 11 September 2007 could be systematically different from already existing bloggers. To address these possible effects, the authors varied the first three regression models to include only those bloggers who started blogging before September 2007. Also, eventual participants are counted as such for all the month beginning from September 2007, i.e. not only beginning with their month of enrollment. Altogether, six regression models were estimated.

The results in table 3 show that the program effect on content popularity of participants for the first three models, where the enrollment dates are taken as breakpoint, range between 7.0% to 7,8%. Eventual participants have an average percentage of higher post popularity of 22% compared to nonparticipants, based on the first OLS model. Taking September 2007 as breakpoint, that systematic effect diminishes to 15.8%. This is, however, justifiable since the program effect of the last three models with September 2007 as breakpoint are higher, i.e. 9.6%, 9.3% and 13%, respectively, implying that many participants may have already started blogging more popular content in preparation of program enrollment, an effect that has not been captured as a program effect within the first three models, resulting in a higher value for the systematic difference between participants and nonparticipants within the first OLS model. For the rest of their analyses, the authors decide to take September 2007 as breakpoint, using the fixed-effect regression model with and without two-stage least-square as their regression framework.

2) Shift in content topics

Based on talking to several frequent Chinese bloggers, the authors identified the stock market, salacious content and celebrities as the most-mentioned topics. Two research assistants then independently classified all the 150 most popular tags in each month to either one of these three categories or a fourth category termed ‘others’, resolving any discrepancies in a meeting. In analogy to the procedure described above, the authors then classified the blog posts in one or more of the four categories and calculated the percentage of each of the three topics for each blogger in each month. Similar to the results about content popularity and as reported in table 2 in the paper, the mean for eventual participants is not only higher for eventual participants compared to nonparticipants, i.e. roughly 3%-5% for the topics stocks and celebrity and 5%-8% for salacious content, but the gap between the two means also gets larger in the test and open application period. This suggests that participants increase their share of posts related to these three categories, compared to nonparticipants. The three calculated topic variables were then taken each as dependent variable in a fixed-effect with and without two-stage least-square regression model, i.e. six models were built. The regression results as displayed in table 5 show that in all three domains, there is a significant shift of content topics of eventual participants, totaling to 6.6% when taking the FE/2SLS models.

3) Shift in content quality

As for content quality, the authors establish four measures: the average percentage of readers' bookmarks for each bloggers' blog post for each month and the average number of characters, pictures and videos of each bloggers' blog posts for each month. While the former should reflect the readers’ satisfaction derived from reading the blog posts, the latter three should measure the blogger’s effort, arguing for a relationship between higher effort and higher quality. As for the other variables, summary statistics as well as comparison of means statistics of these four new variables are also provided in table 1 and table 2 of the paper. Except for the average percentage of number of videos, which is quite low for both groups, and that of the number of pictures, where eventual participants only exceed nonparticipants on average once the open application period has started, the average percentage of bookmarks and number of characters is higher for participants as compared to nonparticipants. Also, the gap between the two means tends to increase over the three periods in favor of the eventual participant group, suggesting that eventual participants are motivated to increase their content quality. This increase in their quality measures can also be seen in table 6, where the authors display the results of their fixed effects with and without two-stage least-square regression models, building two models for each content quality dependent variable.

<u>4) Program effects across participants and over time</u>

Up until now, the authors investigated average program effects on bloggers' behavior. To find out whether these effects vary across different participants, they first create two groups of eventual participants, i.e. those below and those above the threshold of 3000 page views as of August 2007, the month before program announcement. The threshold results from the necessary page views, i.e. 700 for four consecutive weeks being rounded up to 3000, in order to be eligible for the program later on in the open application period. 1,984 (73%) of the 2,722 eventual participants joining the blogging site before program announcement have more than 3000 page views in August 2007. The idea behind the grouping of eventual participants is that bloggers below the threshold - in attempt to increase their chances of being eligible for the program- might shift their blogging behavior more than those above the threshold. The regression formula is adapted to include these two groups of eventual participants; the dependent variables from above, i.e. the popularity, the three content topics and the four quality measures, are each chosen as dependent variable in fixed-effect with and without two-stage least-square regression models. The regression results show that the below-threshold group shifts their topics more relative to the above group. Except for the number of pictures, they also increase their content quality more than the other participants. As the authors intuit, the results suggest that participants who are on or near the threshold of program eligibility change their blogging behavior more than the above-threshold group.

Next, the authors investigate whether the program effects vary over time. The idea is to find out more about the long-term impact and persistence of the program effects. They introduce a new independent variable measuring the number of months since a participant has enrolled in the program. Adapting the regression formula, they re-run the fixed effect with and without two-staged least-square regression models for each of the eight popularity, content topic and quality measures as dependent variable. The results in table 7 in the paper show that no diminishing effect of program impact can be observed for the popularity and content topic measures. As for the latter, the content topic of stock market, where a minimal increase is observed, might be an exception. When it comes to the content quality measures, the results show that the change to higher quality content even increases over time, suggesting that the learning curve of participants is ongoing and sustainable.

After presenting their study results, as summarized in 1) to 4), the authors explore the robustness of their models. First, they further investigate their instrumental variables approach to address the self-selection problem. The idea is to support the exogeneity of the instrumental variables, which is an underlying assumption in the instrumental variables approach employed. Therefore, the authors use a combination of three complementary approaches: propensity-score matching matches participants and nonparticipants based on observables, aiming to correct for a selection bias (DiPrete and Gangl 2004); the Rosenbaum bounds deal with the main assumption of propensity-score matching, i.e that observable characteristics totally account for the selection of bloggers, by estimating how strong the influence of an unobservable must be to completely nullify the results of propensity-score matching (Rosenbaum 2002); the AET-SSS approach then shows, based on the assumption that the strength of the selection on observables and on unobservables is 50:50, whether there still exist positive and significant program effects, excluding thus the possibility that an unobservable variable with an impact of less than 50% would change the positive results (Altonji et al. 2005). Since the authors find that an unobservable variable must influence the selection process by at least 50% and still have positive results in the AET-SSS approach, unobserved variables are unlikely to overturn the positive effect measured via propensity-score matching, which strongly supports the exogeneity of the two instrumental variables.

Second, the authors perform a pretrend analysis, showing that the patterns between participants and nonparticipants do not diverge significantly prior to the program announcement. This in turn increases the confidence in the causal relationship of the program effects on blogging behavior. Third, they re-run their models with different definitions of popular tags. The idea is to counter the possibility that tag popularity and content popularity and topics might be endogenously associated with each other. Furthermore, they aim to control for the possibility that participants might strategically manipulate their tags by attaching more or more popular tags to their blog posts. Taking different measures to redefine popular tags, the authors claim to achieve similar results. Fourth, they remove 133 bloggers considered as outliers, i.e. those whose average monthly posts exceed four standard deviations above the mean, since they may disproportionally influence what is popular or not. Doing so, the authors again claim to achieve similar results. After having explored robustness, the authors summarize their results and discuss their potential welfare implications, ending with possible limitations of the study as well as potential implications for future theoretical studies.

B) Specific contribution and key implications for academia and industry.

It must be noted that this paper largely stands alone in addressing its particular topic, as there appears to be, as the authors themselves point out, not prior literature on its specific topic, at least from an empirical approach, as well as no later literature. According to authors, this is driven by the difficulty of proving causality in such a case (p. 2315) . There is some foundation for their study in the theoretical literature, however, from academics who claim that ad-sponsored business models create the incentive to cater to popular tastes in order to maximize their appeal (Hamilton 2004, McChesney 2004). The impact of this, from a cultural perspective, is claimed to be rather dire by other researchers such as Postman (2005), who holds that the relationship between advertising and media is a destructive force against culture as amusement is prioritized above all else. From the economic side, there has been some recent theoretical work on how ad revenue can incentivize media producers to put out mainly popular pieces, while also duplicating the successful work of competitors, particularly with regard to television programs (Anderson and Gabszewicz, 2006). Ultimately, the paper by Sun and Zhu is breaking rather new ground, but perhaps with limited future potential in its current form, as shall be discussed shortly.

The article is focused specifically on the effects of pay-per-impression ad schemes with revenue sharing on blogger behavior in China. It finds that such schemes do influence the number of "popular" (as defined by the authors) blog posts by participants, with such popular content increasing by roughly 13% for participants relative to non-participants (p. 2322). Furthermore, the effect is shown to maintain itself over time. It is also argued that the "quality" of blog posts sees a relative increase for participants as well, although, as will be discussed shortly, the metrics for quality employed in this article are somewhat contentious. The authors properly put forward the disclaimer that the paper's findings cannot be assumed to transfer to different types of ad revenue models (e.g. pay-per-click or targeted advertising), and the authors admit that the findings may not cleanly transfer outside of China, at least in terms of specific topics of popularity, due to content restrictions in this country (p. 2329).

The key implication for academia that stems from this paper is that ad revenue sharing schemes can indeed have an impact on content curation for participating users of the relevant platform. This is an important assertion, as it supports the idea that content creation may indeed be adapted, in such a situation, to maximize the creator's share of advertising revenue. While far from an absurd idea, it does bring up the important question of whether and/or how content creation might be steered for the purpose of maximizing such revenue in particular situations. Clearly, this would be something worthy of consideration not just for academics, but also firms seeking to plan their marketing activities as efficiently as possible. Unfortunately, the fact should be recognized that since the paper was written, randomized, pay-per-impression ad schemes have in many cases been superseded by targeted and/or pay-per-click models as technology has improved, particularly via such large players as Google's AdSense. Given the great potential for targeted/pay-per-click advertising to more thoroughly deliver relevant ads to relevant segments of consumers, and to verify their reception via click-through rates, the enduring practical relevance of this paper's findings for future use in academia and industry may be rather limited. There are likely still sites and/or markets that employ untargeted, pay-per-impression ads, however. And, on the other hand, the paper does serve as a truly unique and interesting case study of a topic that could be ripe for further research incorporating the latest internet trends and technologies.

C) Overall assessment

Overall, the paper succeeds in studying the phenomenon that it proposes at the onset: the effect, on popularity and quality of content, of sharing ad revenue from un-targeted, pay-per-impression ads with bloggers on a mainstream Chinese portal website. While the methodology is largely sound, and the scope rather rigorously adhered to by the authors, criticisms can be made, particularly with regard to certain choices of qualification and presentation. Specific criticisms can be made on the following points: 1) the definition ascribed to the concept of “quality” in the context of blog posts; 2) some concern about the primary methodology utilized to determine the “popularity” of blog posts. Minor suggestions will also be put forward on small details within tables and the presentation in the text of a specific singular case in such a way that the reader may interpret it as a general rule.

D) Detailed comments

Perhaps the biggest qualitative issue to be taken with the paper is the authors’ approach to the definition of the term “quality” itself, in the context of blog posts, and how they have determined to measure it. The quality measures constructed by the authors are arguably insufficient to capture such a subjective aspect. Number of characters, number of pictures, and number of videos are put forward, along with percent bookmarked, as directly correlated with the nebulous concept of “quality” (p. 2323). While perhaps in many cases these figures are higher in quality posts, higher values for these are not necessary or even sufficient conditions for defining a quality blog post. The authors claim that these metrics are directly related to the effort required to create a post, a relationship that is far from certain given the possibility for short, but well-researched and deeply thought-out blog posts presenting a topic in a much more thorough and arguably "higher quality" manner. Similarly, the logical leap in claiming that effort translates directly to quality, while often the case, is very far from a general rule. It can be allowed that these metrics may hold up in many cases, but they are too imperfect to be making any scientific claims about improvements in blog post "quality" per se. All of this gets to the heart of the difficulty of determining “quality” at an unequivocal level, which is likely beyond the scope of modern philosophy, let alone an article about behaviors around online advertising models. A more apt, but perhaps less appealing, approach would have been to identify the increasing written length and use of media associated with participation in the ad-revenue program as simply this, rather than attempting to label it as "quality" improvements. A more accurate description would likely be something along the lines of “blog post length and media richness”.

In terms of the attribution of the “popularity” of blog posts, there is a lesser criticism to be made. While the method of determining the popularity of blog posts via whether they have received one of the most popular writer-generated tags for a given month is not unsound, it does leave the determination of a topic in the hands of writers who may indeed, as the authors acknowledge (p. 2328), be under the incentive to maximize the popularity of their posts. The concern is that strategic manipulation of tags, i.e. tagging a post with a number of potentially unrelated topics in an effort to maximize page views, could be very easily undertaken. Admittedly, the authors claim to have checked for this through a method for assessing the topic of a blog post via the most commonly used words in a given post (p. 2328), and received “similar” but unreported results. It seems that such a text scanning method is a technique far less subject to manipulation on the part of bloggers, as it would be far harder for them to manipulate the most commonly used words in a post than to quickly and arbitrarily assign a wide range of popular tags once a post is completed. There may have been time or processing-power constraints that dissuaded the authors from employing the post-scanning technique throughout, but these are not presented as limiting factors in the article, which leads the reader to wonder why the technique wasn’t used from the beginning and exactly how “similar” the results really are.

Despite these criticisms, the paper must be praised for its many strengths. Chief among those is its robustness testing. The authors test for a wide range of issues in the data, from the potential impact of unobservable variables to the risk of autocorrelation, all of which seem to back up the robustness of the results (p. 2326-2327). Furthermore, the choice of methodologies for approaching issues related to the problem of participants in the ad revenue sharing scheme self-selecting were particularly well employed, via propensity score matching analysis and the prudent utilization of instrumental variables. And in general, the authors must be praised for selecting a nearly ideal case study for their chosen topic of research, and for staying within its scope throughout.

E) Minor comments

A first small criticism is that the authors chose to discard outlier values only during the robustness testing. While this decision, according to the authors, made no significant difference in the results here either way (p. 2328), there is no clear reason given for not doing so before undertaking their analysis.

There is some concern about the accuracy of the data in Table 2 (p. 2320) with regard to the bookmarking rate of blog posts by non-participants. While it must be stated up front that the data very well could be accurate, the presence of 0 values for non-participants and the striking nature of the disparity between them and participants would in this case warrant some commentary from the authors. If correct, the maximum possible average bookmarking rate for non-participants, accounting for the need to round down to 0.0%, would be 0.04%. Meanwhile, that for eventual participants stands at 1.9% before announcement of the revenue sharing program and 11.2% after open enrollment in the program. This represents at least a 47 times higher likelihood before the program, and at least a 280 times higher likelihood after, for an eventual participant to have a post bookmarked, relative to a non-participant. This is an enormous difference, and is likely even far greater assuming the bookmarking rate is likely less than the maximum possible of 0.04%, as shown in the table, and assuming that the data is in fact accurate. One possibility is that certain high-performing bloggers are hugely pulling up the numbers, and this would add reasoning to the suggestion that outliers in the data should have been discarded before calculations rather than as part of robustness testing. Some greater explanation of this data is likely warranted, as without it the reader is left to wonder about either the accuracy of the data as presented, or the enormous scale of the disparity.

The final comment, which is far more minor still, regards the representation in the text of findings around the potential for the popularity of blog posts to increase for a given blogger along with the duration of participation in the ad revenue sharing program. While the authors point out, as supported by the data, that the program's impact on the popularity of posts written by participants cannot be proven to diminish over time (p. 2325), they continue in the next sentence to say, "To the contrary, for example, the shift toward the stock market [topic] increases with the number of months in the program." While being somewhat of a nitpick, it must be stated that the sentence gives the impression that increases in posts on stock market topics over time is indicative, i.e. an "example", of typical behavior in general. As Table 8 (p. 2326) shows, however, every topic aside from stock markets, including 'general popularity', cannot be seen to have undergone any significant change in posting rate over time. While this is in line with the authors' primary claim that there is no evidence of diminishing effects, it is somewhat misleading to present the singular instance of an increase as an example of anything.

Works Cited

Altonji JG, Elder TE, Taber CR (2005) Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools. *Journal of Political Economics* 113(1): 151–184.

Anderson SP, Gabszewicz JJ (2006) The media and advertising: A tale of two-sided markets. Ginsburgh VA, Throsby D, eds. *Handbook of the Economics of Art and Culture*, Vol. 1, Chap. 18 (Elsevier B.V., Amsterdam), 568–614.

DiPrete T, Gangl M(2004) Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments. *Sociological Method*. 24(1):271–310.

Hamilton JT (2004) *All the News That's Fit to Sell* (Princeton University Press, Princeton, NJ).

McChesney RW (2004) *The Problem of the Media: U.S. Communication Politics in the Twenty First Century* (Monthly Review Press, New York).

Postman N (2005) *Amusing Ourselves to Death: Public Discourse in the Age of Show Business* (Penguin, New York).

Rosenbaum PR (2002) *Observational Studies*, 2nd ed. (Springer, New York).