Managing User Generated Content: A Dynamic Rational Expectations Equilibrium Approach

Dae-Yong Ahn[*] Jason A. Duan[†] Carl F. Mela[‡][§]

June 18, 2014

[*] Assistant Professor, College of Business and Economics, Chung-Ang University, 221, Heukseok-Dong, Dongjak-Gu, Seoul 156-776, Korea; email: daeyongahn@cau.ac.kr; phone: 82 2 820 5944.

[†] Assistant Professor, McCombs School of Business, University of Texas at Austin, One University Station B6700, Austin, Texas 78712; email: duanj@mccombs.utexas.edu; phone: 512 232 8323.

[‡] T. Austin Finch Foundation Professor of Business Administration, Fuqua School of Business, Duke University, 100 Fuqua Drive, Durham, North Carolina, 27708; email: mela@duke.edu; phone: 919 660 7767.

[§] The authors would like to thank Joel Huber, Wagner Kamakura, Vineet Kumar, Oded Netzer, and seminar participants at Carnegie-Melon University, Erasmus University, King Carlos III University, Tilburg University, the University of Chicago, University of Texas, and the 2011 QME Conference for their comments. The authors also thank the NET Institute (www.NETinst.org) for partial financial support of the project.

Abstract

This paper considers the creation and consumption of content on user generated content platforms, e.g., reviews, articles, chat, videos, etc. On these platforms, users' expectations regarding the amount and timing of participation by others becomes germane to their own involvement levels. Accordingly, we develop a dynamic rational expectations equilibrium model of joint consumption and generation of information. We estimate the model on a novel data set from a large Internet forum site and offer recommendations regarding strategies of managing sponsored content and content quality. We find sponsoring content can be effective at creating a self-sustaining network, but once the network tips, sponsored content does little to increase usage.

Keywords: user generated content, marketing strategy, rational expectations, approximate aggregation.
JEL Classification: D71, D83, D84, M15, M31.

1 Introduction

By dramatically lowering the cost of content dissemination and consumption, online communication platforms have engendered a rapid proliferation in global user engagement. Evidence is afforded by a recent ranking done by Google's Ad Planner, listing several user sites with substantial user generated content among the top 20 most trafficked web sites (`YouTube.com`, `Wikipedia.com`, `Mozilla.com`, `Wordpress.com`, `Ask.com`, `Amazon.com` and `Taobao.com`).[1] Coincident with this increase, advertisers are spending 37% more on social media and user generated content sites (UGC), exceeding $4BB annually, or more than 10% of firms online advertising expenditures (eMarketer 2013).

UGC platforms rely upon two behaviors, consuming content (e.g., listening or reading) and generating content (e.g., discussing or writing). Content consumption generates utility via the pleasure of reading or the utility of information. As more content is generated by others, the likelihood of finding content of interest grows. Content generation, like posting video game "cheats" and TV show reviews, yields utility from the reputation effect of being influential, knowledgeable or popular, suggesting utility increases as more content is consumed (Bughin 2007; Hennig-Thurau et al. 2004; Moe and Schweidel 2012; Nardi et al. 2004; Nov 2007).[2] Accordingly, the content generation decision is predicated on beliefs about the number of other people consuming and generating content. As such, users' beliefs about others' participation on the platform are central to the problem of one's own participation, content generation and consumption. In spite of this few, if any papers, explicitly consider the potential role of these beliefs on the growth of UGC networks.

Moreover, expectations about the state of the network can have dynamic implications (Shriver et al. 2013). Because consumption begets generation (and the converse), site participation becomes more likely as aggregate beliefs about the network size increase. As content is durable (i.e., viewers can find older content to be relevant), the current period content generation is analogous to an investment that can generate future returns. Content generation can be costly in terms of time to create it; as a result it induces a tradeoff between

[1] `http://www.google.com/adplanner/static/top1000/`

[2] In this paper we use content and posting interchangeably in which case posting implies the posting of user generated content.

the marginal cost of creating content now intended for future consumption or creating that content later. Though we find evidence of this type of strategic behavior in our data, we are unaware of research that explicitly incorporates this intertemporal trade-off.

We address these points by developing a dynamic, rational expectations equilibrium model of content generation and consumption in the context of heterogeneous users. This rational expectations equilibrium forms the basis of a joint model of site participation, content consumption and generation. Owing to its structural orientation, this approach enables us to address various policy questions of interest to UGC platforms as follows:

- Network Expansion. To increase consumers' utility of consumption, platforms can provide more site sponsored content (SSC); for example, an online forum site could invite experts to create additional content to supplement that of users. On the one hand, increased sponsored content attracts more users who will likely to post more content, because of the increased availability of information. In this instance, sponsored content is a strategic complement to user content. On the other hand, sponsored content can dissuade users from posting content, because sponsored and user content are substitutes from the reader's point of view. The optimal amount of sponsored content, therefore, becomes a question of the relative magnitude of these various effects. In our context, we find that sponsored and user content are strategic complements at low levels of sponsored content, but become substitutes as the sponsored content crowds the user content.

- Network Contraction. Of central interest to network formation is the concept of a tipping point, wherein the platform has a sufficient amount of content to attract readers and a sufficient number of readers to attract content in a self-sustaining manner (that is, the critical mass to become self-sustaining). Without sufficient reading mass, content generators might believe that there is little value in creating content or participating on the site, thereby causing the network to become ensnared in an undesirable equilibrium with very low level of activity for the hosting platform; often referred to as network failure in the economics literature (Liebowitz and Margolis, 1994). We consider several factors that can influence the tipping point of the network:

– User Generated Content. We consider the threshold at which the network contracts (from its self-sustaining level), or alternatively tips (from its low activity level). First, we find that 10.7% of the content level of the self-sustaining UGC network is sufficient to tip the network – if the network focuses upon its most active participants. Marketing strategies such as advertising or incentives and rewards for posting, are not uncommon approaches to attract such users. Second, some network members actively post and read, while others primarily read but rarely, if ever, post, a behavior often called "lurking" (Preece et al., 2004). We find it is not possible to tip a network with just lurkers, although they play an important indirect role in tipping the network; the content threshold for active participants needed to tip the network increases to 16% when the number of lurkers halves. This outcome obtains because lurkers increase overall reading and thus lead to higher posting utility.

– Initial Sponsored Content. An alternative option to tip the network is to substitute SSC for UGC; that is, sponsor content in an effort to attract posting and reading. Two strategies exist to achieve this outcome. First, a firm can seek to jump start the network by sponsoring content early on, and then cease once the network self-sustains. We call this the initial SSC strategy. Such an approach can minimize expenses, as the cost of creating content is only borne by the firm early in the life of the network. We find that this strategy requires an initial amount of SSC at 9.3% of the content level of the self-sustaining network to tip it.[3] Alternatively, a firm can use a regular SSC strategy by sponsoring posts at a steady level and thus change users' beliefs about steady state content. We find that this strategy requires a regular amount of SSC at 8% of the content level of the self-sustaining network to tip it. The tipping point is lower than a transient increase because user expectations about increased future network size affect current participation decisions.

[3] Note the sponsored content required to tip the network is lower than the user content needed to tip the network. As we discuss shortly, the difference arises when there is diminishing marginal returns to the utility of user posting.

- Content Quality. In addition to exploring the number of users, it is also possible to assess the effects of the quality of posts they read. We consider two such potential manipulations on the part of the platform: changing the quality of the site sponsored content and changing the quality of the user content.

  - Changing Sponsored Content Quality. For both mature and nascent networks, site sponsored content can have higher average quality than UGC. In our counterfactuals, we consider the implications of higher quality posting strategies by the firm and find that tipping points can be substantially lowered. For the initial SSC strategy, higher quality can lower the amount of SSC required for tipping from 9.3% to 1%; for the regular SSC strategy, the amount of SSC required for tipping can be lowered from 8% to 1%. Higher quality posts can also be especially effective at growing mature networks.
  - Changing User Content Quality. Sites can also filter user posts by making them harder to access or removing them altogether by moderating them. We consider two such filters. The first involves eliminating low quality posts. The second, motivated by concerns over fair use, involves removing high quality posts under the presumption that misappropriation of intellectual property tends to involve higher quality information. We find it is optimal to filter a small amount of low quality content, on the order of 1% to 1.25% for this particular site. Though deleting high quality content makes the site worse off in traffic, the effects are slight if small amounts are deleted.

Our model, owing to its structural nature and its ability to capture rational expectations equilibrium, is the first to our knowledge to shed light upon the role of these different strategies in network tipping and traffic in the context of UGC. Overall, results suggest that tipping is quite feasible with a relatively small amount of high quality sponsored content. This strategy may be quite cost effective as the expense in creating content manifests only in the early stages of the network. Examples of web sites that have pursued this strategy include `Soulrider.com` (Shriver et al. 2013), a wind surfing site which jump started the network by inviting experienced surfers to generate high quality content in its infancy. Of course, to grow the network, this site could have alternatively provided a smaller and more regular stream of high-quality sponsored content, or instead targeted regular users with incentives to join the network, and our model yields insights regarding the potential of these various approaches to build the network.

Also of note, the results of our policy experiments are profoundly affected when dynamics or the beliefs about the participation of others is not allowed to evolve with changes in the system as is common in descriptive research. For example, we find the estimated amount of UGC to tip the network increases from 10.7% to 19.0% of the self-sustaining equilibrium content when beliefs are ignored – an error of nearly 80%. The amount of content needed to tip the network is overestimated in the absence of modeling expectations, because users are not allowed to update their beliefs about potential increases in user content in future periods. These results indicate the dynamic rational expectations equilibrium approach we develop is critical when assessing how user generated content is affected by firm strategy and changes in the environment. In sum, by integrating beliefs regarding the effect of others' consumption and generation of content on one's own content decisions with a rational expectations equilibrium, we develop a model that enables us to explore the growth of UGC network. Our approach is quite general and applies to many content generation and consumption contexts ranging from chat rooms to journal publications to video sharing sites (where users post and consume content), we estimate this model using a proprietary data from a web site where users generate and consume content in the form of Internet forums (e.g., `tv.com/forums/`, `espn.go.com/nfl/forums`, `city-data.com/forum/`, `birdforum.net`, `petforums.com`, `archerytalk.com`, etc.).[4]

In the next section, we elaborate upon how our model of site participation, content creation and content consumption differs from prior literature on social networking in general, and UGC in particular. We then discuss our data and context and use this information to construct our model. Then we explore some of the theoretical properties of our model, discuss identification and estimation, detail our results and conduct policy simulations.

[4] It is worth noting that, in our application, the number of people reading a post is not reported and the UGC is not rated. As discussed in the conclusions section, it is possible to extend our framework to accommodate these contextual variations.

2 Literature Review

Our work is related to the growing empirical literature in marketing on social networking and interaction (e.g., Ansari et al. 2011; Stephen and Toubia 2010; Bulte 2007; Hartmann 2010; Nair et al. 2010; Katona et al. 2011; Iyengar et al. 2011). Our work deviates from the social networking literature inasmuch as we consider user sites with large numbers of agents such that any single agent's participation is not likely to have a sizable effect on aggregate content consumption or generation. To exemplify this point, consider a user who posts a review on an ESPN.com forum; this agent might focus more upon the sizable number of interested viewers consuming their content than any given viewer who consumed it. In such instances, it becomes feasible to model the dynamic social engagement choices of agents in a structural fashion, because we do not need to condition on the behavior of all other individual agents (e.g., Hartmann 2010) but only on the aggregate states such as the total number of posts or reads.

Likewise, our research is related to the burgeoning literature on UGC (Albuquerque et al. 2010; Chevalier and Mayzlin 2006; Dellarocas 2006; Duan et al. 2008; Shriver et al. 2013; Ghose and Han 2011; Moe and Schweidel 2012; Zhang et al. 2012) that considers the joint behavior of content consumption and generation.[5] Our research extends this work in two ways.

First, we allow users' expectations about aggregate site engagement to change with the state of the network; i.e, we consider a rational expectations equilibrium. This is material because changes in expectations regarding the number of users contributing, for example, can affect whether agents visit a site, consume, or write. It is therefore necessary for any policy intervention to accommodate potential changes in expectations, such as network tipping. Second, we consider dynamic decision making. UGC, like advertising, decays in efficacy over time, analogous to the advertiser problem outlined in Dubé et al. (2005). Because content is somewhat durable, those who post consider not only whether there posts are read now, but also in the future. This sets up a trade-off between costly current period investment in

[5] Ghose and Han (2011) consider a dynamic structural model of mobile phone content usage based on consumer learning; they do not jointly model the dynamics in content consumption and generation. Our work is also complementary to theirs inasmuch as the dynamics in our model reflect expectations about future readership for posts rather than uncertainty in the usage experience.

the stock of posts for future reading against the costly future investment. Intuitively, one might expect that lower future costs will encourage users to shift content generation to later periods, but this must also be weighed against the loss of current period readership.

These two advances lead us to develop a dynamic structural model given rational expectations of UGC. As a dynamic structural model, our paper is similar to Huang et al. (2011) who consider the blogging behavior of the employees of an IT firm. An important point of difference is that we use a rational expectations equilibrium framework to link individual behavior to aggregate state transitions. In addition, our work extends the structural literature by considering how the quality of site content affects overall user engagement. As a result, we can conduct policy analyses on the role of site content quality on user engagement.

The solution to dynamic problem in the context of a large UGC network involves each user forming beliefs about many thousands of other users; a task that is both computationally infeasible for the researcher and cognitively unwieldy for a UGC site user. To address these concerns, we extend the approximate aggregation approach of Lee and Wolpin (2006) and Krusell and Smith (1998). In this approach, users reason that the aggregate evolution of the network should be consistent with sum of decisions made by all the members of the network, thereby enabling users to form beliefs about aggregate state transitions in lieu of each individual's states. As a result, the aggregate state transitions across all the users can vary with changes in the primitives of the system, yielding a structural interpretation of the social engagement problem.

Though we draw upon the approximate aggregation approach, our work fundamentally differs from past instantiations in many respects. First, a single unit of supply (a post) can be consumed (read) by many users at the same time. In past research on labor or capital, a single unit of supply is consumed by a single agent. As such, we have no market clearing condition. Rather, equilibrium arises from a balance of different network effects, such as competition for readers (direct network effects) and the attraction of readers (indirect network effects). Second, because content is generated and consumed by a single agent, our problem differs considerably from previous markets wherein producers and suppliers differ. Third, we adapt the concept to an entirely new context, UGC networks.

Finally, because we consider strategies by which the UGC platforms can become self-sustaining, our work is related to the literature on tipping (Dubé et al. 2010; Katz and Shapiro 1994). In that research, tipping is defined as "the degree of market share concentration due to indirect network effects" (p. 216, Dubé et al., 2010). In our context, the indirect network effects for the platform arise from generation and consumption rather than software and hardware; tipping is defined as achieving the critical mass for the network to become self-sustaining. In addition, we also consider user heterogeneity through segmentation.

In sum, our contribution is to develop a dynamic structural model of site participation, content generation and content consumption for a large number of users and use this model to evaluate network effects and draw implications regarding how the site which hosts these interactions should manage the volume of its content.

3 Data

We overview the context we model as a prelude to model development and estimation. Our data come from a large Internet site devoted to a common interest, which includes a forum where persons can discuss various topics much like fans would discuss a sports team, its players or various games. However, our model is not pathological to these data but can be applied more generally to a number of UGC contexts.

3.1 Data Context

Figure 1 outlines the data context and the decisions we observe; content consumption, content generation and site participation. Users *consume content* (such as reading reviews about a video game) generated by others for their interest in information. An increase in content generation can lead to an increase in content consumption, because users are more likely to find information of interest (Stigler 1961).

An increase in content consumption can also lead to an increase in *content generation;* those who post content presumably do so, because they are motivated to have their posts read by others (Bughin 2007; Hennig-Thurau et al. 2004, Moe and Schweidel 2012; Nardi et al. 2004; Nov 2007). There is also a potential direct network effect of content generation on content generation; as more content appears, competition for readers increases.

Rational users base their own *participation* decisions (whether they visit the site) on

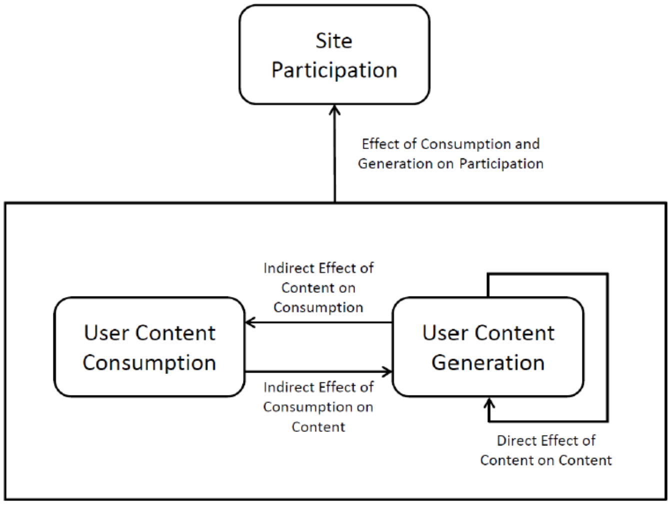


Figure 1: Model Overview

activity of others (e.g, Katz and Shapiro 1998; Ryan and Tucker 2012; Dubé et al. 2010). More participation in aggregate leads to more content and higher reading rates, thus leading to greater individual participation utility. When the mass of participation becomes sufficiently high, there is a threshold beyond which a network can become self-sustaining and below which the network implodes. Our paper extends prior research by modeling *both* the network participation decisions and the content generation/consumption decisions.

Finally, the durability of posts in our data can create incentives for users to be forward-looking in managing them. Specifically, users can create content in the current period for consumption in the future instead of creating content in the future for consumption in the future. This induces a tradeoff between current and future costs of posting and between current and future readership.

Below, we provide some descriptive statistics to characterize the three decisions captured in our data as well as the potential for strategic content management by users.

3.2 Descriptive Statistics

We collect two months of forum participation data in user log files from October through November 2009 and use this as our basis of exploration for social engagement. User log files include the complete visit, read and post history for each registrant. We consider total reads and posts by each user on a daily basis. This yields 19,461,572 user-day observations.

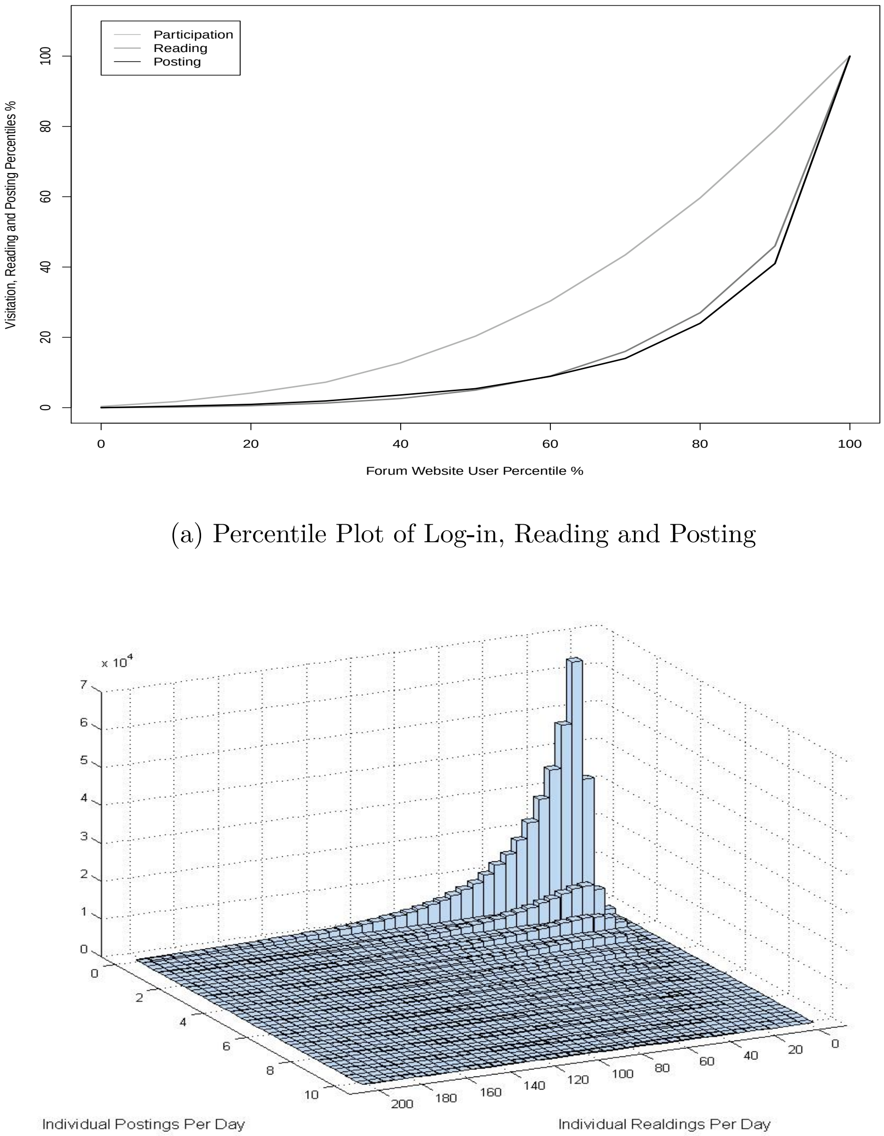
Variable	Mean	Std. Dev.	Min.	Max.
Site Participation (n_{it})	0.42	0.49	0.00	1.00
Forum Reading (r_{it})	17.97	47.42	0.00	345.00
Forum Posting (a_{it})	0.42	1.53	0.00	19
Individual Post Stock (k_{it})	1.19	2.97	0.00	19.32

Table 1: Descriptive Statistics

Table 1 reports the descriptive statistics for the key variables (excluding 0.05% of outliers) used in our analysis. The table indicates participation is frequent, with 42% of users visiting the site on any given day. Forum reading is far more prevalent than forum posting, and there is significant variation in forum reading and posting across individuals as indicated by large standard deviations of these variables. The average individual post stock equal to 1.19 is quite low, but some users are heavily invested in the site with larger post stocks.

Further considering the differences across users, we present the distribution of site engagement, defined as participation rate, reading rate and posting rate per user. From Figure 2a, we note that the observed rates of reading and posting are remarkably close to the endemic "80/20" rule observed in many marketing contexts, 20% of the users are responsible for 76% of posting and 73% of reading. This observation again suggests the need to accommodate unobserved heterogeneity in reading and posting.

Finally, Figure 2b plots the joint distribution of content generation and consumption, conditional on site participation. The figure shows that reading is more common than posting, as there is a substantial percentage of users who read more than 100 posts a day, but very few users create more than 5 posts a day. Users' reading and posting rates are highly correlated, as users with higher posting rates tend to have higher reading rates. These observations further underscore the need to accommodate unobserved heterogeneity jointly in reading and posting. Given site participation, all users read posts, but some do not create posts. Therefore, the reading model should accommodate an interior solution for the utility optimization problem, whereas one can characterize content generation via a discrete choice model with the option of choosing zero content.



(a) Percentile Plot of Log-in, Reading and Posting

(b) Joint Distribution of Reading and Posting

Figure 2: Summary Plots of Reading and Posting

3.3 Exploratory Analysis

To explore the potential for indirect network effects and dynamics characterized in Section 3.1, we conduct a regression analyses using a random data sample of activities of 600 users over 61 days.

3.3.1 Individual Reading and Aggregate Post stock

In this subsection we regress the daily amount of reading of individuals against aggregate UGC stock (unit in 1,000 postings) and individual post stock using a random data sample of activities of 600 users over 61 days. Table 2 shows that higher aggregate UGC stock indeed leads to higher daily amount of individual reading, whereas individual post stock does not significantly affect daily reading, consistent with the reading utility model. This suggests that readers are affected by the amount of content on the site.

Variable	Parameter Estimate	*t*-value	*p*-value
Aggregate UGC Stock	0.079*	3.72	0.00
Individual Post Stock	0.11	1.32	0.19
Weekend Effect	−1.03*	−2.03	0.04

Table 2: The Effect of Aggregate UGC stock on Individual Reading

3.3.2 Individual Posting and Aggregate Reading Rates

Next, we explore the effects of the population's average rate of reading per post on the individual postings. If users obtain utility from others who read their posts, higher reading rate should increase the utility of posting. In addition, since lagged higher individual post stock decreases the marginal utility of posting, we expect to see negative correlation between lagged individual post stock and current postings. Because daily individual postings are small integers, we fit a generalized linear model (GLM) using the Poisson family.

Table 3 reports the results of this regression. The results suggest a significant positive effect of the average rate of reading per post and a significant negative effect of the lagged individual post stock on the likelihood of posting. The aggregate UGC stock does not have a direct significant effect on individual posting conditional on the average reading per post, consistent with a process wherein posting utility is mediated by reading.

Variable	Parameter Estimate	z-value	p-value
Average Reading per Post	0.057*	2.15	0.037
Aggregate UGC Stock	0.012	1.40	0.16
Lagged Individual Post Stock	−0.0065*	−2.26	0.024
Weekend Effect	−0.063*	−2.79	0.0053

Table 3: The Effect of Average Reading Rate per Post on Individual Posting

3.3.3 Suggestive Evidence of Dynamic Behavior

Following the approach outlined in Chintagunta et al. (2012), we conduct exploratory analyses of dynamic behavior. Evidence in Chintagunta et al. (2012) relies on the autocorrelation between future states and current behaviors, implying that users consider future states in current decisions. In our context, content generation costs appear higher on the weekend, as evidenced by the large decrease in posts observed then (perhaps because of the increased opportunity costs of leisure time and lower Internet access). Hence, if content generation is higher (lower) the day right before the weekend (weekday), this postie correlation is consistent with strategic management of posts. Likewise, as agents can create costly content today or tomorrow for consumption tomorrow, there may be an incentive to delay content creation to when future consumption rates are higher. We analyze these timing decisions by fitting via GLM using the Poisson family.[6] The results in Table 4 demonstrate a negative effect (p-value < 0.05) of the reading per post of the next period on current period postings, suggesting that users may delay content generation until there is an audience available to read them. A similar effect is evidenced for the impeding weekend by the indicator of the day before weekend (weekday) (p-value < 0.01), suggesting that users may move content generation before weekend when the costs are higher.[7]

[6] We caution that the following analyses are suggestive evidence as there are not formal statistical tests of dynamics (Manski 1993; Magnac and Thesmar 2002). The relationship to the dynamics in our model is motivated in Section 4.7 and Web Appendix D.3.

[7] For a robustness check, we also test three additional models by eliminating the effect of weekend/weekday next day (Model 2), the weekend effect (Model 3) and by adding the reading per post of the second next period (Model 4). The results show the same pattern of correlations. Overall, this exploratory regression analysis is consistent with the content generation model.

Variables	Model 1			Model 2		
	Estimate	z-value	p-value	Estimate	z-value	p-value
Average Reading per Post (current period)	0.14*	3.96	7.41×10^{-5}	0.092*	2.72	0.0065
Average Reading per Post (next period)	0.053*	-2.18	0.029	-0.059*	-2.40	0.016
Aggregate UGC Stock	4.85×10^{-7}	0.68	0.50	1.23×10^{-6}	1.79	0.073
Lagged Individual Post Stock	-0.013*	-4.19	2.69×10^{-5}	-0.013*	-4.33	1.51×10^{-5}
Weekend Effect	-0.051*	-2.15	0.032	-0.42	-1.78	0.075
Day Before Weekend/Weekday	0.11*	4.18	2.82×10^{-5}	-	-	-
	Model 3			Model 4		
Average Reading per Post (current period)	0.13*	5.43	5.65×10^{-8}	0.088*	2.55	0.011
Average Reading per Post (next period)	0.051*	-2.15	0.031	-0.0031	0.11	0.91
Average Reading per Post (2nd next period)	-	-	-	-0.11*	-3.70	0.00021
Aggregate UGC Stock	5.39×10^{-7}	0.77	0.44	1.55×10^{-6}∗	2.03	0.042
Lagged Individual Post Stock	-0.013*	-4.18	2.89×10^{-6}	-0.0077*	-2.63	0.0086
Day Before Weekend/Weekday	0.11∗	4.55	5.28×10^{-6}	-	-	-

Table 4: The Effects of Future Average Reading Rate per Post and Weekend/Weekday on Individual Posting

Collectively, Tables 2 through 4 suggest the potential for indirect network effects and dynamics to affect network formation and growth in the context of UGC. We consider these possibilities more formally in the modeling section, discussed next.

4 Model

Figure 1 in Section 3.1 outlines the decisions we seek to model. First, we consider content consumption in the face of heterogeneous quality of UGC in Section 4.1. Second, we outline the role of consumption on content generation in Section 4.2. Third, we discuss site participation conditional on these other two decisions in Section 4.4. Finally, we discuss users' strategic behaviors in content generation implied by the dynamic model in Section 4.7.

In sum, we consider i) a population of M users' decisions (indexed by $i = 1, \ldots, M$) to participate on a content sharing web site, $n_{it} = \{0, 1\}$, at period t ($t = 1, \ldots, T$) and, ii) conditional on participation ($n_{it} = 1$), how much content to consume (read), r_{it}, and iii) how much content to generate, a_{it}. Users choose these three actions $\{n_{it}, r_{it}, a_{it}\}$ to maximize their utility conditional on their expectations regarding overall participation of others in the network. Though the decision to participate on the site is made first, it is contingent on expectations on the utility of reading and posting obtained after visiting. Hence, we solve this problem via backward induction by first presenting the content consumption and generation models before the participation model.

4.1 Content Consumption

4.1.1 Reading Utility

Readers consume content when benefit exceeds costs. The utility of reading is incumbent upon the total content available, because an increase in the number of others' posts enhances the likelihood that a user finds items of interest. To formalize this notion, our model uses order statistics for post quality which is analogous to what Stigler (1961) shows in the derivation of minimal price given the number of price searches. Let readers face a distribution of the entire stock of posts, denoted by K_t, whose qualities, denoted as $Q_1, \dots, Q_{K_t}$, are distributed iid $Uniform\,[L, U]$.[8]

Individuals read the posts of the highest quality. Let the qualities be ranked as their order statistics $Q_{[1]} \leq Q_{[2]} \leq \cdots \leq Q_{[K_t]}$. Each order statistic, $Q_{[k]}$, has the distribution:

$$Q_{[k]} \sim \frac{K_t!}{(k-1)!\,(K_t-k)!} \left(\frac{Q_{[k]}-L}{U-L}\right)^{k-1} \left(\frac{U-Q_{[k]}}{U-L}\right)^{K_t-k} \frac{1}{U-L}, \tag{1}$$

which is a linear transformation of the Beta distribution (i.e., $(Q_{[k]}-L)/\,(U-L)$ has a $Beta(k, K_t+1-k)$ distribution). So the expected quality

$$E\left(Q_{[k]}|K_t\right) = (U-L)\,\frac{k}{K_t+1} + L. \tag{2}$$

If individual i reads the r_{it} highest quality postings, the expected utility is

$$u^R\left(r_{it}\right) = E\left(\sum_{k=K_t-r_{it}+1}^{K_t} Q_{[k]}\right) = \frac{(U-L)}{K_t+1}\left[\left(K_t+\frac{1}{2}\right)r_{it} - \frac{r_{it}^2}{2}\right] + Lr_i. \tag{3}$$

Reparametrizing $\alpha_1 = U$ and $\alpha_2 = U - L$, one can define

$$u^R\left(r_{it}\right) = \frac{\alpha_1 K_t + (\alpha_1-\alpha_2) + \frac{1}{2}\alpha_2}{K_t+1} r_{it} - \frac{\alpha_2}{K_t+1}\frac{r_{it}^2}{2}. \tag{4}$$

The utility of reading can be further simplified when K_t is a large number using the approximation $K_t + 1 \approx K_t$ and $\left[K_t + (\alpha_1-\alpha_2)/\alpha_1 + \alpha_2/2\alpha_1\right]/\,(K_t+1) \approx 1$. Thus reader i's

[8] In Section 7 we generalize our analysis to consider allowing site sponsored content to differ in its quality from user generated content.

utility for reading becomes

$$u^R(r_{it}) = \alpha_1 r_{it} - \frac{\alpha_2 r_{it}^2}{2K_t}. \tag{5}$$

The reading utility is higher when $\alpha_1 = U$, which represents the upper limit on perceived content quality, is higher and lower when $\alpha_2 = U - L$, which represents the uncertainty of the content quality, is higher. Given α_1 and α_2, the marginal utility of reading a post, is increasing with posts K_t. The result follows intuitively from a greater likelihood of finding content of interest when there are more posts. It also provides a microeconomic foundation for empirical regularities detailed in recent research (Ransbotham et al., 2012). In sum, this utility evidences diminishing marginal returns from reading at a decreasing rate in the total number of posts and higher quality.

It is not necessary that all the users have the same ordering of post quality nor the same quality distribution itself. We can introduce individual heterogeneity parameter, ζ_i, in the reading utility

$$u^R(r_{it}) = (\alpha_1 - \zeta_i) r_{it} - \frac{\alpha_2 r_{it}^2}{2K_t}, \tag{6}$$

which implies that the perceived average content quality, $\alpha_1 - \alpha_2/2 - \zeta_i$, is heterogeneous.

4.1.2 Reading Costs

Next we consider the cost of reading. We assume the cost has a quadratic form that reflects an increasing scarcity of time or attention as more items are read, so that

$$c^R(r_{it}) = \kappa_{1it} r_{it} + \kappa_{2i} \frac{r_{it}^2}{2}, \tag{7}$$

where implies a convex cost function if κ_{1it} and κ_{2i} are both positive. Cyclicality, such as the weekend effect, is accommodated by allowing κ_{1it} in equation (7) to vary over time, i.e., $\kappa_{1it} = \kappa_{1i} w_t$ where w_t is a weekend indicator. The cost parameters κ_{1i} and κ_{2i} are heterogeneous across users.

The users' total payoff from reading is therefore expressed as utility less cost, or

$$u^R(r_{it}) - c^R(r_{it}) = (\alpha_1 - \zeta_i - \kappa_{1it}) r_{it} - \left[\frac{\alpha_2}{K_t} + \kappa_{2i}\right] \frac{r_{it}^2}{2}, \tag{8}$$

Given this utility, the expected optimal amount of reading, r_{it}^*, is solved from the first order condition (FOC),

$$r_{it}^* = \frac{\alpha_1 - \zeta_i - \kappa_{1it}}{\alpha_2/K_t + \kappa_{2i}}. \tag{9}$$

Due to heterogeneity in reading costs across segments, r_{it}^* differs across segments. After user i decides to visit the web site, she realized a contextual shock, ν_{it}, which is not observed by the econometrician. We assume that the observed amount of reading by i is r_{it}^* multiplied by individual specific random shock (ν_{it}) so that $r_{it} = r_{it}^*\nu_{it}$.[9] Therefore, ex post amount of reading will also differ across users in the same segment. As ν_{it} is realized after the user's site participation decision, the user's ex ante decision to visit the site at time t depends only on the ex ante expected optimal amount of reading defined by equation (9). In sum, a reader's optimal level of consumption increases with the overall level of content on the site and the quality of posts.

4.2 Content Generation

4.2.1 The Per-Period Utility of UGC Generation

Site users derive utility from others reading their posts. The average rate of reading per post based on rational expectations is used to model the reading likelihood, because a user on our site cannot observe the exact amount of reading for each of her posts (there is no counter for the "number of views" in our data).[10] This expected rate of reading per post y_t is defined by

$$y_t = \frac{R_t}{K_t} = \frac{\sum_{i=1}^{M} E(n_{it}r_{it}^*|K_t, \zeta_i)}{K_t}. \tag{10}$$

Equation (10) shows two competing effects of the aggregate UGC stock K_t on y_t. First, there is a primary demand effect of K_t in the numerator as the expected total amount of reading increases with the supply of content. This constitutes an indirect network effect on posting from reading. Second, there is a competitive effect of K_t in the denominator as more postings will dilute the reading rate per post. This constitutes a direct network effect of posting on posting. Therefore, the net effect K_t on y_t can be positive or negative.

[9] Because ν_{it} is realized after a user's decision on whether she visits the content site, ν_{it} is independent of the site participation decision and uncorrelated with K_t. It is not imperative to impose any parametric distribution on ν_{it}, although we assume ν_{it} to be exponential in Appendix C.1 to facilitate maximum likelihood estimation.

[10] In Appendix A, we show that the users' expected reading rate per post y_t can be closely approximated by the observed amount of reading per post under the assumption of rational expectations when the number of users and the UGC stock K_t are both very large - so actual reading rates can be used in our model estimation. However, it is necessary to recompute this rational expectation equilibrium in our counterfactual analyses (See Section 4.6 for more details).

Following the advertising literature, we assume that posted information follows a geometric decay over time with a parameter ρ (Clarke 1976; Mela et al. 1997; Dubé et al. 2005). In our case, the decay rate is exogenous and relates to obsolescence. For example, posts about basketball games from preceding weeks are less relevant than similar posts from preceding days. The aggregate stock of posts therefore has the following form:

$$K_t = \sum_{\tau=0}^{t} \rho^{t-\tau} A_\tau = \rho K_{t-1} + A_t \tag{11}$$

where $\rho < 1$ is the discount rate, and A_t is the number of new posts in period t. Likewise, let the individual stock of posts and new posts at t by user i be k_{it} and a_{it}, where

$$k_{it} = \sum_{\tau=0}^{t} \rho^{t-\tau} a_{i\tau} = \rho k_{i,t-1} + a_{it}. \tag{12}$$

Given that users form rational expectations for the reading rate y_t, the per period utility from generating content using the constant relative risk aversion (CRRA) utility function with diminishing marginal return is

$$u^P(a_{it}) = u^P\left((\rho k_{i,t-1} + a_{it})\, y_t\right) = \frac{\left((\rho k_{i,t-1} + a_{it})\, y_t\right)^{1-\gamma}}{1-\gamma}. \tag{13}$$

4.2.2 Costs of UGC Generation

Content is not costless to create (Ghosh and McAfee, 2011). These costs include the time to draft and post the content, weighed against the opportunity costs of other activities. For a sample of our data, the mean number of words per post is 72.2 with a standard deviation of 108.8. Karat et al. (1999) report that the average typing rate for composition is 19 words per minute, meaning that creating content takes about 4 minutes on average and can range up to 15 minutes. With multiple posts evident in our data, content generation can take upwards of an hour. Given that opportunity costs of time differ during a week and that Internet access may as well, it is likely these costs vary exogenously with the day of the week.

The cost of posting is specified as

$$c^P_{it}(a_{it}) = (\tau_{it} + \xi_i)\, a_{it} - \varepsilon_{it}(a_{it}), \tag{14}$$

where the random error in the cost function, $\varepsilon_{it}(a_{it})$, has a generalized extreme value (GEV) distribution. The time-invariant component of the linear marginal cost ξ_i, which is allowed to be heterogeneous across users, is assumed to have a latent segment model. In addition, τ_{it} models cyclical effect such as a weekend effect, $\tau_{it} = \tau_i w_t$, where w_t is the weekend indicator. We also assume the cyclical effect τ_i to be idiosyncratic for different latent segments. The heterogeneity in the cost of posting captures the potential unobserved individual-level differences in posting.

4.2.3 Optimal UGC Generation

We presume a user chooses the number of postings a_{it} (amount of content to generate) that maximizes the discounted expected sum of per-period utilities minus per-period costs to obtain the following value function

$$V_i\left(s_{it}, \varepsilon_{it}\right) = \max_{a_{it}, a_{i,t+1}, \cdots} E\left\{\sum_{k=t}^{\infty}\left[u^P(a_{ik}) - c_{ik}^P(a_{ik})\right]\right\}. \tag{15}$$

In this dynamic optimization problem, $s_{it} = \{k_{i,t-1}, K_t, \tau_{it}\}$ and ε_{it} are the state variables and the number of per-period postings a_{it} is the control variable. Posting by users a_{it} is treated as a discrete variable, because over 99% of users post only from zero to ten posts a day. Hence, the posting decision a_{it} is a discrete choice of number of postings, i.e., $a_{it} \in A = \{0, 1, 2, \ldots, \bar{a}\}$,

The value function of this optimization problem has the Bellman's equation

$$V_i(s_{it}, \varepsilon_{it}) = \max_{a_{it} \in A}\{u^P(a_{it}) - \bar{c}_{it}^P(a_{it}) + \varepsilon_{it}(a_{it}) + \beta E[V_i(s_{i,t+1}, \varepsilon_{i,t+1})|s_{it}, a_{it}]\}, \tag{16}$$

where the $\bar{c}_{it}^P\left(a_{it}\right) = \left(\tau_{it} + \xi_i\right) a_{it}$ represents the non-random part of posting costs.

Define the integrated value function $\tilde{EV}_i(s_{it}, a_{it})$ as

$$\tilde{EV}_i(s_{it}, a_{it}) \;\; = \;\; \int_{s_{i,t+1}} \int_{\varepsilon_{i,t+1}} V_i(s_{i,t+1}, \varepsilon_{i,t+1}) p(s_{i,t+1}, \varepsilon_{i,t+1}|s_{it}, \varepsilon_{it}, a_{it}) ds_{i,t+1} d\varepsilon_{i,t+1}.$$

We can derive the probability of writing a_{it} content postings conditional on site participation as

$$P(a_{it}|s_{it}, n_{it} = 1) = \frac{\exp(u^P(a_{it}) - \bar{c}_{it}^P(a_{it}) + \beta \tilde{EV}_i(s_{it}, a_{it}))}{\sum_{a'_{it} \in A} \exp(u^P(a'_{it}) - \bar{c}_{it}^P(a'_{it}) + \beta \tilde{EV}_i(s_{it}, a'_{it}))}. \tag{17}$$

4.3 Modeling Heterogeneity in UGC Consumption and Generation

We model the combined heterogeneity parameters in the reading and posting in a joint distribution with J latent segments,

$$[\zeta_i, \kappa_{1i,}\kappa_{2i}, \xi_i, \tau_i] \sim \sum_{j=1}^{J} p_j I\left(\zeta_i = \bar{\zeta}_j, \kappa_{1i} = \bar{\kappa}_{1j}, \kappa_{2i} = \bar{\kappa}_{2j}, \xi = \bar{\xi}_j, \tau_i = \bar{\tau}_j\right). \quad (18)$$

In the formula above, $\bar{\zeta}_j$, $\bar{\kappa}_{1j}$, $\bar{\kappa}_{2j}$, $\bar{\xi}_j$ and $\bar{\tau}_j$ are the segment-specific values for the parameters in the content consumption and generation models.

As reading and posting costs follow a distribution that is to be jointly estimated, our model captures a wide array of interdependent behaviors between reading and posting. If, for example, one segment has a low posting cost and a low reading cost and another segment has a high posting cost and a high reading cost (reflective of the tendency of some individuals to read and post more than others), then reading and posting will be correlated across users.

In the reading model represented by equation (9), it is obvious that the data cannot separately identify α_1 and $\bar{\zeta}_j$, $j = 1, \ldots, J$. Hence, we constrain $\bar{\zeta}_j$ such that $\sum_{j=1}^{J} \bar{\zeta}_j = 0$.

4.4 Site Participation

Prior to posting and reading, a user must decide whether to visit the UGC web site and this decision is predicated upon the net expected utility from consuming and generating content should the user decide to visit. Hence, the utility from visiting the site ($n_{it} = 1$) includes utilities from expected posting and expected reading,

$$\begin{aligned} u^V(n_{it} = 1) &= \mu_1 E \max_{r_{it}} [u^R(r_{it}) - c^R_{it}(r_{it})] + \\ &\quad E \max_{a_{it}} [u^P(a_{it}) - c^P_{it}(a_{it}) + \beta \tilde{EV}_i(s_{it}, a_{it})] + \eta \varepsilon_{it}(n_{it} = 1) \end{aligned} \quad (19)$$

where μ_1 and η are scale parameter that receptively rescale the utility of reading and the contextual shock relative to the utility of posting. The contextual shock $\varepsilon_{it}\,(n_{it} = 1)$ represents the exogenous cost for a user to visit the site at period t, and it is assumed to be known to the user but not the econometrician.

The corresponding utility from not visiting the site ($n_{it} = 0$) contains three components. First, users continue to obtain utility from others' reading previous post stock, given by

$u^P(\rho k_{i,t-1}y_t) + \beta \tilde{EV}_i(s_{it}, a_{it} = 0)$. Second, μ_{0j} is a segment-specific intercept and captures the utility from time spent on alternative pursuits when one does not visit the site. Third, there is a random shock $\varepsilon_{it}\,(n_{it} = 0)$. Therefore, the utility from not visiting the site is obtained by summing these three components:

$$u^V(n_{it} = 0) = \mu_{0j} + u^P(k_{i,t-1}y_t) + \beta \tilde{EV}_i(s_{it}, a_{it} = 0) + \eta\varepsilon_{it}(n_{it} = 0). \tag{20}$$

A user chooses to visit the web site if $u^V(n_{it} = 1) > u^V(n_{it} = 0)$ and vice versa.

We assume $\varepsilon_{it}\,(a_{it})$, $\varepsilon_{it}\,(n_{it} = 0)$ and $\varepsilon_{it}\,(n_{it} = 1)$ have iid Type-1 Extreme Value (Gumbel) distributions, resulting in a nested logit model of site participation and content generation given site participation.

Let the inclusive value of posting content be

$$IV_{it} = \ln \sum_{a_{it}\in A} \exp(u^P(a_{it}) - \bar{c}^P_{it}(a_{it}) + \beta \tilde{EV}_i(s_{it}, a_{it})), \tag{21}$$

then we derive the choice probability of visiting the site as

$$P\left(n_{it} = 1|s_{it}\right) = \tag{22}$$

$$\frac{\exp\left\{\mu_1 E \max_{r_{it}}[u^R(r_{it}) - c^R_{it}(r_{it})] + \eta IV_{it}\right\}}{\exp\left\{\mu_{0j} + \eta\left[u^P(k_{i,t-1}y_t) + \beta \tilde{EV}_i\left(s_{it}, a_{it} = 0\right)\right])\right\} + \exp\left\{\mu_1 E \max_{r_{it}}[u^R(r_{it}) - c^R_{it}(r_{it})] + \eta IV_{it}\right\}}$$

and $P(n_{it} = 0|s_{it}) = 1 - P(n_{it} = 1|s_{it})$.

4.5 State Transitions

The state transitions are as follows. First, the random shocks, ε_{it}, are assumed to be i.i.d. over time and across individuals and independent of the other state variables in s_{it}. Second, the individual stock k_{it} evolves deterministically $k_{it} = \rho k_{i,t-1} + a_{it}$. Third, the day of week effects, w_t evolves deterministically over time. Lastly, the aggregate UGC stock, K_t, is defined as $K_t = \sum_{i=1}^{M} k_{it}$, and hence it evolves deterministically given K_{t-1} and a_{it}; $i = 1, \ldots, M$. However, we assume that from the perspective of any individual user, K_t evolves stochastically given K_{t-1}, but independent of their own action a_{it}. When the site has a very large number of users, every user i will neither perfectly observe the actions $a_{i't}$ and stocks $k_{i't}$ of all the other users nor believe her own action a_{it} has any influence on the aggregate UGC stock, K_t. This claim is similar to the assumption in perfect competition where no agents in the market believe their own output can change the total supply. If we impose the rational expectations constraint, then user i's belief about the state transition for K_t must coincide with the actual behavior by all users on the site. This will be discussed in detail next.

4.6 Rational Expectations and Approximate Aggregation

Rational expectations require that users' beliefs about K_t be consistent with its actual transition, which is the sum of all individuals' posting decisions. This observation becomes critically important in policy simulations, because the evolution of K_t is neither exogenous nor invariant to a change in policy that might affect users' participation levels. Users' beliefs can change in response to a change in the strategy of the site.

4.6.1 Approximate Aggregation

Extending an approximate aggregation approach to the rational expectations equilibrium pioneered by Krusell and Smith (1998), we first formulate an individual's beliefs on how the aggregate state variable K_t evolves over time as follows

$$K_t = \omega_0^K + \omega_1^K K_{t-1} + \omega_2^K w_t + \varepsilon_t^K, \tag{23}$$

where w_t is the weekend indicator and ω_2^K represents cyclical effect. The parameters ω_0^K, ω_1^K, ω_2^K for the stock of the aggregate content are determined by the rational expectations equilibrium.

We posit a first degree order of the lag in the state transitions to be consistent with the primitives in the consumer model to ensure that the approximate beliefs regarding the aggregate state transitions are consistent with the Markovian structure in the underlying individual posting model.[11] From an individual's perspective, there is a degree of uncertainty

[11] Note that the order of the state transition equations cannot be higher than the order of the individual level model, else the individual level model would fail to account for consumer's beliefs about these higher order states. Here we assume individuals only use one lagged K_{t-1} to predict K_t and hence it implies an AR(1) model for K_t. Conceivably, individuals may use more than one lagged stock to predict K_t.Were they to use an AR(q) model then $K_{t-2}, \ldots, K_{t-q}$ would also have to be in the set of state variables in the dynamic optimization problem. As a surfeit of state variables can induce computational dimensionality constraints, the most parsimonious state transition model possible for K_t is desirable from a computational perspective. In Section 3.3, we test and find the AR(1) model is best model for our data.

about the evolution of K_t; we express this uncertainty using ε_t^K, which is a zero-mean random error given K_{t-1}.

Our model also assumes that individual users approximate the expected average reading rate per post as a function of K_t with equation

$$y_t = \omega_0^y + \omega_1^y K_t + \omega_2^y w_t, \tag{24}$$

where ω_2^y is again for the effect for weekend. Equation (24) approximates equation (10) which does not have a closed form for the function y_t of K_t. When the number of users is very large, the observed quantity of average reading per post can closely approximates the expected one.[12] See Web Appendix A for details. The parameters ω_0^y, ω_1^y, ω_2^y are also determined by the rational expectations equilibrium.

4.6.2 Consistency of Approximate Aggregation

In reality, K_t is deterministic given the actions of all individuals

$$K_t = \rho K_{t-1} + \sum_{i=1}^{M} a_i \left(k_{it}, K_t, \tau_{it}, \varepsilon_{it}\right). \tag{25}$$

Using equation (25) directly to calculate users' rational expectations requires us to assume each user knows all the other users' policy functions $a_i \left(k_{it}, K_t, \tau_{it}, \varepsilon_{it}\right)$ as well as their post stocks k_{it}. Complete knowledge of the behavior of thousands of other users is an unrealistic assumption, which imposes a large informational burden on every individual user. In addition, this assumption places $k_{it}, i = 1, \ldots, M$ in every user's set of state variables, causing the "curse of dimensionality", which renders the dynamic programming problem intractable.

On the other hand, approximate aggregation (assuming bounded rationality) only requires that K_t and y_t predicted by equations (24) and (25) coincide with the real K_t and y_t. This implies that agents need only be able to form rational beliefs regarding the transitions of the aggregate states. Using an initial guess for the parameters ω_0^K, ω_1^K, ω_2^K and ω_0^y, ω_1^y, ω_2^y, we compute individual behaviors n_{it}, a_{it} and r_{it}^*. Aggregating across persons, we recompute K_t and y_t and recompute individual behaviors, iterating back and forth between

[12] We aslo test approximating R_t in equation (10) with $R_t = \omega_0^R + \omega_1^R K_t + \omega_2^R K_t^2 + \omega_3^R w_t \cdot K_t$, from which we derive the approximation for $y_t = R_t / K_t$ in estimation. We find the estimation results remain unchanged, so the approximation by equation (24) is robust.

the individual-level and aggregate models until convergence. Web Appendix B details the algorithm used to simulate a rational expectations equilibrium. The parameters $\omega_0^K,\ \omega_1^K, \omega_2^K$ and $\omega_0^y,\ \omega_1^y, \omega_2^y$ are re-estimated in every step of the iterations to find the fixed point of the rational expectations equilibrium.[13] In sum, the use of approximate aggregation enables us to accommodate heterogeneity in a rational expectations equilibrium model.

We explore some theoretical properties of our the rational expectations equilibrium with approximate aggregation by simulation in Web Appendix D.

4.7 Dynamic Strategic Behavior

In this final subsection, we afford some intuition regarding the nature of dynamics implied by our model. The dynamics in our model rests on two integrated foundations: i) the individual intertemporal substitution of content creation and ii) the expected indirect network effects.

We begin by discussing the individual's intertemporal substitution of content. Considering a simplified and stylized version of the UGC generation problem where we treat the discrete posting decision, a_t, as a continuous variable, costs as linear, and abstract away from the error (we suppress individual index i for clarify)

$$V(s_t) = \max_{a_t \in A}\{(\rho k_{t-1} + a_t)^{1-\gamma} / (1-\gamma) - \tau_t a_t + \beta V(s_{t+1})\}. \tag{26}$$

We can derive the Euler equation from this simplified problem as follows:

$$\frac{(\rho k_{t-1} + a_t)^{-\gamma} y_t^{1-\gamma} - \tau_t}{(\rho k_t + a_{t+1})^{-\gamma} y_{t+1}^{1-\gamma} - \tau_{t+1}} = \beta\rho. \tag{27}$$

Equation (27) captures the tradeoff between creating content for the next period in the current period against creating the content for the next period in the next period. If one chooses to accelerate content creation from the next period to the current period, they gain i) the utility from those that read that content this current period $[y_t(\rho k_{t-1} + a_t)]^{-\gamma}$, and ii) the cost difference of creating that accelerated content this current period (τ_t) rather than next period (τ_{t+1}), but iii) lose some utility from others reading tomorrow because of the decay in content and the discounted utility $\beta\rho$. Overall, this expression implies dynamics

[13] In estimation, the aggregate states are observed (reflecting the current equilibrium), so no iteration to re-estimate $\omega_0^K, \omega_1^K, \omega_2^K$ and $\omega_0^y, \omega_1^y, \omega_2^y$ is necessary. In the counterfactual analyses, states need to be computed.

involve shifting content creation to align with periods with higher reading rate y_t and lower posting cost τ_t. Moreover when $\rho = 0$, there is no inter-temporal tradeoff of generating content, which is the case when content is not durable (one cannot create content today for consumption tomorrow). Finally, we note that all else equal, an increase in entering stock, k_{t-1} implies an attendant reduction in a_t to maintain the equality. The rational behind this relationship is that there is an optimal level of current and past content which can be obtained either by having a larger stock of old content or an increase in current content. In Web Appendix D.3, we show how the descriptive analysis in Section 3.3.3 is consistent with intertemporal conditions implied by equation (27). This provide some additional empirical support of dynamics in our model.

Next, we consider the role of expected network effects in dynamics as discussed in Section 4.6. Although the Euler equation in the simplified model outlines the ratio of content across periods, it is not informative about the absolute level of content. Yet beliefs about future content available also affects the overall level of content. In this regard, the individual decision to maintain stock k_t depends not only on the current period network effects (represented by y_t and K_t) but also the expected future network effects as implied by the aggregate state transition equations. This is analogous to a firm that makes investment choices given their expectations on future profitability and interest rates. For example, if a user believes the site manager will continuously sponsor content, as opposed to only once, then the user expects more readers in the future which will affect the generation of current postings.

5 Estimation and Identification

5.1 Estimation

Within each iteration of the likelihood optimization algorithm, an efficient estimation approach using maximum likelihood requires solving both i) the nonlinear dynamic optimization problem for every individual user and ii) the rational expectations equilibrium for the aggregate reading and posting. The computational cost of this approach is therefore considerable as it involves i) iterations for rational expectations within ii) iterations for the fixed point solutions for the dynamic program within iii) iterations for the likelihood routine.

To facilitate the second set of iterations, we design a two-step estimation approach as in Rust (1994), first estimating the state transition equation for the aggregate UGC in (23) and the reading-per-posting as a function of the UGC in (24), and second estimating the parameters in reading, posting and site-participation models.

In the first step of this approach, we estimate the state transition equation for the aggregate UGC in equation (23) and the reading-per-posting as a function of the UGC in equation (24). We obtain the MLE estimates of the regression coefficients $\omega_0^K,\ \omega_1^K, \omega_2^K$ and $\omega_0^y, \omega_1^y, \omega_2^y$, which capture the evolution of the aggregate states in the data under the current equilibrium. Because the observed K_t and y_t reflect the current equilibrium, no iteration is needed to re-estimate $\omega_0^K,\ \omega_1^K, \omega_2^K$ and $\omega_0^y, \omega_1^y, \omega_2^y$. This equilibrium assumption is not likely to hold in our counterfactuals wherein we do need to recompute the expectations.

In the second step, we estimate the structural parameters in the individual reading, posting and site-participation models. In this step, we use the first-step results to estimate the structural parameters. The reading and posting models are estimated jointly with MLE, using the joint likelihood function of reading and posting for each individual user i

$$\left\{\sum_{j=1}^{J} p_j \prod_{t=1}^{T} \text{Exponential}\left(r_{it} | \frac{\alpha_1 - \bar{\zeta}_j - \bar{\kappa}_{1j} w_t}{\alpha_2 / K_t + \bar{\kappa}_{2j}}\right) \frac{\exp\left(u^P(a_{it}) - \bar{c}^P(a_{it}) + \tilde{EV}_j(s_{it}, a_{it})\right)}{\sum\limits_{a'_{it} \in A} \exp\left(u^P(a'_{it}) - \bar{c}^P(a'_{it}) + \tilde{EV}_j(s_{it}, a'_{it})\right)}\right\}. \tag{28}$$

The site-participation model in equation (22) is estimated as a binary choice model with the likelihood function

$$\left\{\sum_{j=1}^{J} p_j \prod_{t=1}^{T} \left[P(n_{it} = 1 | s_{it})^{n_{it}} P(n_{it} = 0 | s_{it})^{1 - n_{it}}\right]\right\}, \tag{29}$$

given the estimated parameters in the reading and posting models. The derivation of these likelihood functions is detailed in Web Appendix C.

To facilitate the first set of iterations pertaining to solving the dynamic programming problem during the second step of the estimation, the estimation algorithm parallels Dubé et al. (2012), which is a maximum likelihood estimator using mathematical programming with equilibrium constraints (MPEC). Su and Judd (2010) show that the two-step pseudo maximum likelihood (2S-PML) estimator discussed above is consistent. We bootstrap to compute the 95% confidence intervals. See Web Appendix C.2 for details.

5.2 Identification

5.2.1 Post Stock Decay Parameter ρ

As indicated in Section 4.2, the information contained in a posting gradually becomes obsolete. We model this phenomenon by imputing an exogenous decay parameter ρ to the post stock in our model. The decay rate ρ is identified and estimated using a secondary data set, which records a sample of posts and their respective histories of how many times they are read over time. In this data set, we observe i) that there is a decline over time in the number of times that a particular posting is accessed by forum users after it is posted and ii) the average reading rate per post, aggregate UGC stock and participation rate are stationary in time. The data are consistent with our modeling assumption that a post has a finite lifetime with a decay parameter ρ. When the average reading rate per post is stationary over time as observed in our data, the decay parameter is identified by the ratio of the times that a post is read in periods $t-1$ and t. Under the exponential decay assumption, this ratio equals the decay rate in the amount of reading per post.[14]

5.2.2 Other Parameters

In the aggregate state transition equations (23) and (24), the coefficients are identified from the time series structure of the aggregate level data – specifically, from the autocorrelation for the K_t and correlation between the y_t and K_t. Next, we consider the identification of the parameters in the reading and posting models from the second stage estimation. Note that the optimal individual level reading level r^*_{it} for person i at time t is equal to $(\alpha_1 - \bar{\zeta}_j - \bar{\kappa}_{1j}w_t)/\left(\alpha_2/K_t + \bar{\kappa}_{2j}\right)$ if person i belongs segment j per equations (9) and (18). This expression suggests that α_1 is not identified, because one can divide the numerator and denominator by any constant and obtain the same ratio. For this reason, we normalize α_1 to 1, which also achieves scale normalization. Heterogeneity in the reading $(\bar{\zeta}_j, \bar{\kappa}_{2j})$ and posting $(\bar{\xi}_j)$ models can be inferred from differences in individuals' mean reading and posting levels over time from the panel structure of the data. To identify the cyclical effect in reading and

[14] We test the exogeneity of the decay parameter by regressing the log ratio of the times that a posting is read between periods $t-1$ and t on the aggregate UGC and the average reading rate per post at period t. For our data (see the results in Section 6.1), we find neither factor to be statistically significant. This suggests that the decay parameter ρ is independent of the aggregate activities of the forum and therefore exogenous.

posting $\bar{\kappa}_{1j}$ and $\bar{\tau}_j$, we classify the days of a week into weekdays and weekends and use a dummy variable that is set 0 for weekdays and 1 for weekends. Hence, these parameters are identified by differences in the mean amount of reading and number of postings between weekdays and weekends. The parameter that captures the diminishing marginal returns for the utility in posting, γ, is identified by the observed difference in mean posting levels at different levels of individual post stock. In general the discount factor is not identified in dynamic discrete choice models (Manski 1993; Magnac and Thesmar 2002). Hence we set $\beta = 0.99$.[15] Finally, there are scale parameters to estimate in the site participation model μ_0, μ_1 and η. Conditional on the parameters estimated above for the reading and posting model, these parameter estimates follow from standard identification arguments for the logit with panel data on site participation.

6 Results

6.1 Decay Parameter and Initialization of Post Stock

As indicated in Section 4.2, the post stock is incumbent upon the decay rate of a post. The exogenous decay parameter ρ is estimated using an auxiliary dataset collected by the Internet site regarding when a sample of users' posts were visited by other users. The decay in the number of users clicking on these posts over time is informative about their durability. We consider a random sample of 473 forum postings in the first week of the sampling period. The dataset records the daily number of times that these posts are read for 20 days. The average number of times that a post is read on the first day is 7.59 ($sd = 13.9$), the second day 4.90 ($sd = 5.4$) and the third day 3.49 ($sd = 3.1$), etc.

The decay parameter is identified by the ratio of the times that a posting is read in periods $t-1$ and t. Let z_{kt} be the number of times post k is read is in the t-th period after it is posted. We estimate ρ using the model $z_{kt} \sim Poisson\,(\rho^{t-1}\lambda_k)$, where λ_k captures the heterogeneity in the amount of reading among posts. The MLE estimation using generalized linear models for $\widehat{\log\rho} = -0.30$ (0.0036), so the decay parameter estimate is 0.74.

As there is no history of posts prior to the initial week, individuals' initial post stocks

[15] We also estimate the model with $\beta = 0.98, 0.95, 0.90$ and our insights remain unchanged.

Model	AR(1) for UGC stock	Average Reading Rate per Post
Intercept ω_0^K or ω_0^y	$2.89 \times 10^4 [0.19, 5.59] \times 10^4$	6.02 $[4.14, 7.90]$
Lagged Aggregate UGC stock ω_1^K	0.93 $[0.86, 0.99]$	–
Current Aggregate UGC stock ω_1^y	–	5.43×10^{-6} $[0.36, 10.5] \times 10^{-6}$
Weekend effect ω_2^K or ω_2^y	-6.92×10^3 $[-8.80, -5.04] \times 10^3$	-0.55 $[-0.69, -0.42]$
Residual R^2	0.89	0.51

Table 5: Estimation Results for Aggregate UGC Stock Transition Equation and Average Reading Rate per Post Equation (with 95% confidence intervals in brackets)

in the first week of the data are unobserved. Hence, the individual post stock is computed by setting the initial stock at zero and recursively applying equation (12) using the 61-day posting data repeatedly until the individual post stock reaches a steady state. The individual's steady state is then re-used as the initial post stock to calculate the individual post stock for the 61-day data. Most of the users in our sample have been using the forum for a long time prior to the sampling period, so their post stocks are likely to have reach the steady state at the starting period of our data. Aggregate stock K_t is computed by aggregating individual stocks.

6.2 Aggregate Variables under Rational Expectations

Section 4.6 outlines the aggregate state transition model that captures the rational expectations process. The estimation results for the AR(1) model in (23) and (24) are reported in Table 5. The results provide evidence of strong auto-correlation ($\omega_1^K = 0.93$) for the aggregate UGC stock. The average reading rate per post is an increasing function of aggregate UGC stock ($\omega_1^y = 5.43 \times 10^{-6}$), which implies a positive network effect of posting. The weekday effects for Monday, Tuesday and Wednesday in model (23) are not significantly different from Sunday, whereas the effects for Thursday, Friday and Saturday are significantly negative, which implies lower posting activity for these days of a week. Because these day effects differ only between the weekday and the weekend, we group them into a single weekend indicator for Thursday through Saturday. The weekend effect in model (24) is significantly negative, which means lower reading activity for these days.

We also test an AR(2) model $K_t = \omega_0^K + \omega_1^K K_{t-1} + \omega_2^K K_{t-2} + \omega_3^K w_t + \varepsilon_t^K$ using the second order lag K_{t-2}. We find the second lag coefficient ω_2^K to be not significant (p-value $= 0.73$).

Durbin-Watson test for the residuals of the AR(1) model $K_t = \omega_0^K + \omega_1^K K_{t-1} + \omega_2^K w_t + \varepsilon_t^K$ has the p-value $= 0.63$, which cannot reject the null hypothesis that the autocorrelation of the residuals is 0. Therefore, our data supports the assumption that users can use approximate aggregation and AR(1) to predict K_t on the basis of rational expectations.

6.3 Content Consumption, Generation and Site Participation

We randomly selected a sample of 600 users to estimate the individual-level model. The amount of reading and number of postings for each individual in the sample are recorded for 61 days from October 1st to November 30th, 2009. If both reading and posting are zero for a user in a certain day, we conclude that the user did not visit the site that day. Table 6 reports the parameter estimates for the model for two segments of users.[16]

We begin by discussing the results for the content generation and consumption models. The two segments of the content generation model are specified to share a common posting utility parameter, γ, in equation (13) but differ with respect to their posting costs $\bar{\xi}_j$ in equation (14), as heterogeneity in both costs and utilities are not separately identified. The two segments of the content consumption model share a common utility parameter α_2, but differ with respect to utility parameter $\bar{\zeta}_j$ and cost parameters $\bar{\kappa}_{1j}$ and $\bar{\kappa}_{2j}$. Note that the estimated segment specific effect $\bar{\zeta}_j$ is equal and opposite in magnitude across two segments; this reflects the normalization needed for identification.

Comparing the two segments, the second one is larger in size and evidences higher reading and posting costs; hence, this group of users read less often and rarely posts content. Thus, we denote the segments "heavy users" and "light users" ("lurkers" by Preece et al. 2004). Also of note, the weekend effect $\bar{\tau}_j$ in the posting cost function is positive, so the users tend to post less on a weekend.

Next, we consider the parameter estimates in the site participation model. As indicated in Table 6, the heavy user segment participates more often because of their higher expected reading and posting utilities. This implies that reading and posting behaviors are correlated across users. Of note, conditional on reading and posting utilities, no parameters differ sig-

[16] We also test three segment of users. However, the BIC for the three-segment model is higher than the two-segment model.

Parameters	First Segment Heavy Users	Second Segment Light Users
Content Generation Model		
Utility coefficient γ	$0.79\ [0.74, 0.85]$	
Cost coefficient $\bar{\xi}_j$	$1.29\ [1.13, 1.47]$	$6.83\ [6.02, 7.12]$
Weekend effect $\bar{\tau}_j$	$8.99\ [1.12, 14.72]$	$9.60\ [2.46, 14.75]$
Content Consumption Model		
Utility coefficient α_2	$0.60\ [0.48, 0.75]$	
Utility heterogeneity coefficient $\bar{\zeta}_j$	$-0.70\ [-0.54, -0.76]$	$0.70\ [0.54, 0.76]$
Weekend cost effect $\bar{\kappa}_{1j}$	$-0.049[-0.11, 0.001]$	$-0.0052\ [-0.039, 0.0075]$
Quadratic cost coefficient $\bar{\kappa}_{2j}$	$0.026[0.021, 0.028]$	$0.011\ [0.008, 0.015]$
Site Participation Model		
Intercept, μ_{0j}	$27.12[1.01, 53.94]$	$4.46\ [0.37, 11.22]$
Reading scale parameter μ_1	$0.51\ [0.031, 1.32]$	
Gumbel scale parameter η	$0.97\ [0.94, 0.99]$	
Heterogeneity		
Segment size	$43\%\ [38\%, 48\%]$	$57\%\ [52\%, 62\%]$

Table 6: Estimation Results for the Parameters in Content Generation, Consumption and Site Participation Models

nificantly between segments in the site participation model. This suggests that heterogeneity in site participation decisions are largely predicated on reading and posting utilities.

To show model-fit, we simulate user reading, posting and vising data given the estimated parameter values in Table 6. We then calculate the mean absolute percentage error (MAPE) between the simulated aggregate UGC, average reading rate per post and number of visitors across 61 observed days and the corresponding observed values. These values are, respectively, 7.59%, 4.47%, 7.04%, which show our model fit the data very well.[17]

7 Managing Content Quality and Quantity

In this section, we investigate how the site can manage its content to improve its traffic. We explore two types of content management strategies: influencing UGC and creating site

[17] We also fit two static models to the data. In the first model, we let users realize the flow utility from the net present value of their posts, but without the inter-temporal ability to choose when to post. We find this static model fits the data much more poorly (log-likelihood = -1.50774×10^5) than our dynamic model (log-likelihood = -9.9972×10^{-4}). In the second model, we aggregate data to the weekly level to filter out daily variation and estimate a static model. As the likelihoods are not comparable between daily and weekly models, we compute MAPE for the weekly average UGC and reading rate and find our model has better predictive performance than the weekly aggregate data model.

sponsored content (SSC). We consider SSC that looks identical to UGC from the perspective of the reader (e.g., the sponsorship is not revealed), except, potentially, in their quality. We focus on this type of content to ensure that the parameters in our UGC reading model are valid for inferring how SSC affects user behavior.

When a site sponsors content, readers choose the best content available across the entire set of UGC and SSC; if SSC is of better average quality than UGC, then it attracts more readers from UGC. To be more precise, define SSC quality distribution parameters $\alpha_1^S = U^S$, $\alpha_2^S = U^S - L^S$ analogous to the definition of the UGC quality distribution. By allowing α_1^U, α_2^U and α_1^S, α_2^S to differ, we can assess the role of user and sponsored content quality on site traffic, content consumption and generation.

The optimal levels of reading r_{it}^S and r_{it}^U for user and sponsored content are the solutions to the optimization problem of the total reading pay-off. In Web Appendix E, we derive the optimal total amount of reading and how it is allocated between UGC and SSC given by

$$\begin{bmatrix} r_{it}^{*S} \\ r_{it}^{*U} \end{bmatrix} = \begin{bmatrix} \left(\frac{\alpha_2^S}{K_t^S} + \kappa_{2i}\right) & \kappa_{2i} \\ \kappa_{2i} & \left(\frac{\alpha_2^U}{K_t^U} + \kappa_{2i}\right) \end{bmatrix}^{-1} \begin{bmatrix} \left(\alpha_1^S - \zeta_i - \kappa_{1it}\right) \\ \left(\alpha_1^U - \zeta_i - \kappa_{1it}\right) \end{bmatrix}. \tag{30}$$

Using this approach, we consider several content management strategies. First, we consider how the addition of SSC will affect current traffic. Second, we explore the role of SSC and UGC quality and quantity in tipping the network. We then conclude by considering how filtering posts can affect network traffic.

7.1 Increasing Site Traffic by SSC

In the context of the site's *current* state of user engagement, the issue of how sponsored content affects site traffic and user engagement is germane. Increased site traffic is material because revenue typically arises from advertising, and advertising revenue increases with visits.

We simulate the effect of quality and quantity of SSC on UGC. Without loss of generality, we manipulate the quality of SSC by altering the lower bound of quality, α_2^S. The upper bound of SSC quality is represented by α_1^S, which we set to equal α_1^U. This approach presumes that the site does not create relatively "bad" content. Notice that higher α_2^S also means that SSC has higher mean quality ($\alpha_1^S - \alpha_2^S/2$) and less quality variation ($\alpha_2^{S^2}/12$) than UGC. We set the quality of SSC to one of two levels: i) equal to the quality of UGC (by letting $\alpha_2^S/\alpha_2^U = 100\%$) or ii) substantially higher than that of UGC ($\alpha_2^S/\alpha_2^U = 10\%$). The resulting aggregate UGC, number of visitors and reading rate per post (normalized to the percentages of the corresponding values in the currently observed data) in equilibrium over a 70-day simulated sequence versus the levels of SSC is plotted in Figure 5.

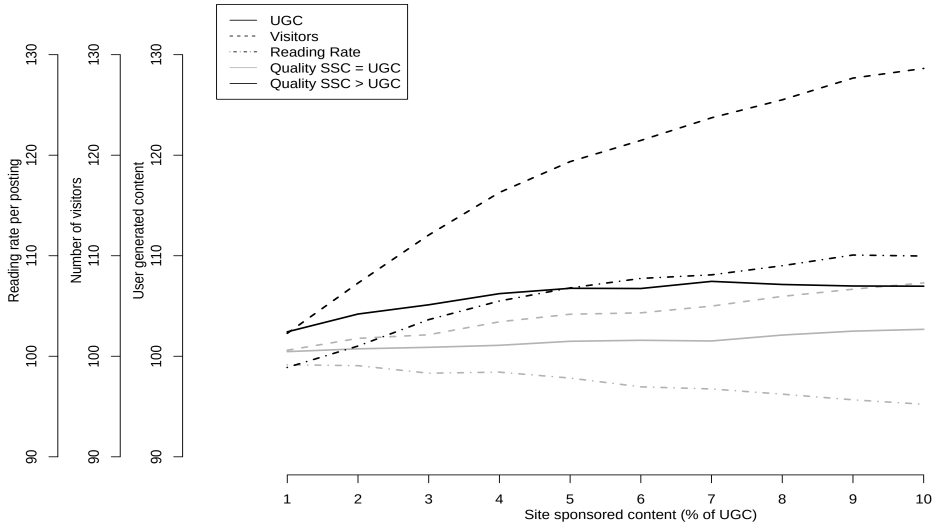


Figure 3: Effect of SSC strategies on increasing site traffic. The horizontal axis represents the amount of SSC as the percentage of the observed aggregate UGC in the data. The vertical axis depicts the percentage changes in user engagement.

Figure 5 demonstrates that SSC increases the number of visitors only slightly when the quality of SSC is equal to that of UGC (by 5% for a 10% increase in overall content). If the site can generate sufficiently large quantity of much higher quality SSC ($\alpha_2^S/\alpha_2^U = 10\%$), the site can increase the participation rate by a much larger 25%. Hence, the sponsored content arc elasticity is about .5 and 2.5 for these respective cases.

The increment in UGC is much less pronounced owing to the competition effect between UGC and SSC, with an arc elasticity of about 0.6 in the high quality case and 0.1 in the low quality case. In the high quality SSC case, the amount SSC from 7% to 10% of the observed UGC dampens user content generation: user posts are strongly affected by the shift of readers to higher quality sponsored posts.

The reading rate per post actually decreases to about 96% in the case when UGC and SSC have equal quality, also because of the competition effect. In the high quality SSC case, the growth in reading rate per post is moderate, a 9% increase.

Overall, we conclude that there is potential to grow the network primarily with very high quality SSC, but that UGC will be adversely affected if the level of SSC exceeds about 7%.

7.2 Network Tipping

In the preceding subsection, we considered a counterfactual analysis predicated on the current equilibrium state observed in the data, wherein the network has already become self-sustaining in terms of high user participation. Another potential equilibrium for the UGC platform is a state of extremely low user participation, where individual postings, reading and site-visiting are all close to zero. This also constitutes a stable equilibrium, because low UGC stock attracts very few readers, which will further attenuate posting activity. The site converges to an equilibrium of extremely low activity, which can be interpreted as the implosion of the network. This kind of network contraction has often been reported in the online community. For example, the UGC site TVTome.com, which used to have a large network of users, has since dwindled and the domain name has been sold to TV.com.

To prevent network contraction, one way to tip the network to a self-sustaining high-activity equilibrium is to attract *user* content to the site (e.g., via marketing or advertising). A second approach, such as the strategy employed by `Soulrider.com`, is to sponsor a sufficient level content to tip the network – either with a large initial amount, or a regular but smaller stream of posts, or both. If it takes a large amount of user posts to tip the network, it might be sensible to “jump start” the network by sponsoring posts rather than relying solely on user posts. It may be better for the site to sponsor posts early on than at a constant rate, because once the network tips, the site will no longer need to bear the expense of sponsoring. While a definitive answer to which of these three strategies (initial user posts, initial sponsored content only and regular sponsored content) depends upon actual costs, one must first understand how these strategies affect network growth and exactly how much content is necessary to tip the network. Quantifying this amount is our aim in this section.

7.2.1 UGC on Tipping

We first consider the role of UGC stock on network tipping. Figure 4 demonstrates the relationship between the initial UGC, defined as the stock at the start of the network, and the subsequent steady-state equilibrium level of forum activities the network will ultimately attain (normalized to the percentages of aggregate UGC, the number of visitors and the reading per post in the self-sustaining high-activity equilibrium). The critical UGC stock needed to tip the network is 10.7%. That is, when the initial UGC is below the 10.7% of the UGC in the high-activity equilibrium, the forum will collapse into the low-activity equilibrium wherein site participation rate reduces to only about 3%, UGC to only about 0.5% and reading per post to about 8%. In contrast, when the initial UGC stock is 10.7% or greater, the site will reach 100% of the high activity.

It is worth noting that modeling rational expectations profoundly affects the tipping point of the network. When beliefs about the participation of others is not allowed to evolve with changes in the system, we find 19% of the current observed levels of UGC is needed to tip the network.

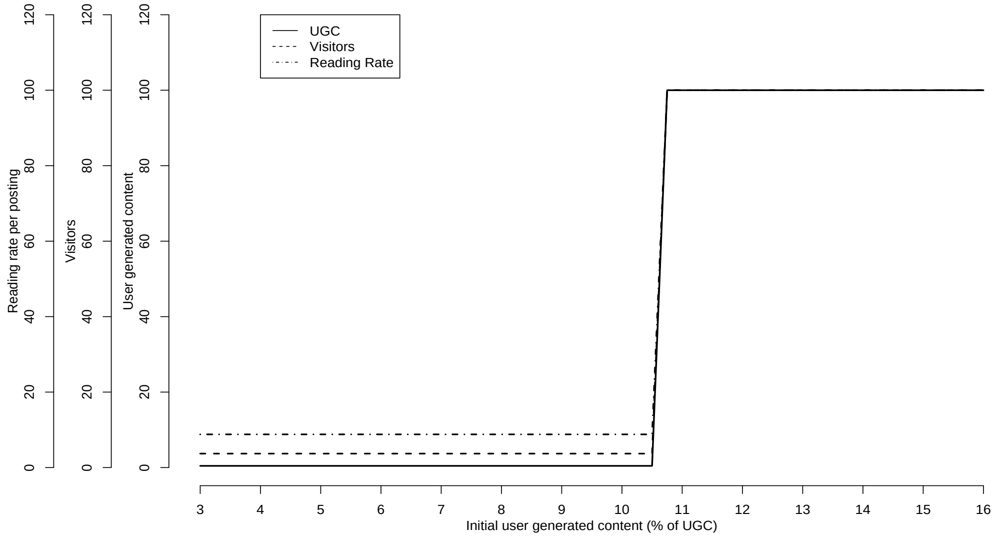


Figure 4: The critical point of initial UGC. The horizontal axis represents the percentage of initial user stock as a fraction of the current average levels of UGC observed in our data. The vertical axis depicts the steady state site usage.

Next, we consider the role of user heterogeneity in tipping. Recall, we observe two segments. The heavy user segment, comprised of 43% of users, tends to both consume and generate content frequently. The light user segment tends predominantly to read. This light user segment is sometimes characterized as the lurking segment (Preece et al. 2004). Because they generate no content, lurkers cannot easily tip the network. However, they can have a considerable indirect effect on tipping, as those in the heavy user segment obtain more utility from content generation because of reading by lurkers. Our simulations find that the critical UGC stock needed for tipping is raised from 10.7% to 16% if the lurker segment is cut in half (though there is little effect if this segment is reduced by only 20%). To our knowledge, this is the first study to quantify the effect of a lurking segment on tipping rates. A key implication is that, if lurkers are cheap to attract, it can be more efficient to grow this segment in order to increase the marginal impact of heavy users on tipping.

Another interesting aspect of heterogeneity pertains to the tipping point of UGC stock required from each segment. Because heavy users generate more content, one might surmise that targeting this group would tip the fastest. Indeed, for the heavy user segment, an initial endowment of user post stock at 11% of the high-activity equilibrium level can tip the network when we set the initial stock of light users to zero. In contrast, for light users, an initial endowment of user post stock at 80% of the high-activity equilibrium level can tip the network when we set the initial stock of heavy users to zero. Hence, the heavy user segment's role in network tipping is far more substantial, but a large population of lurkers can also tip a network.

7.2.2 SSC on Tipping

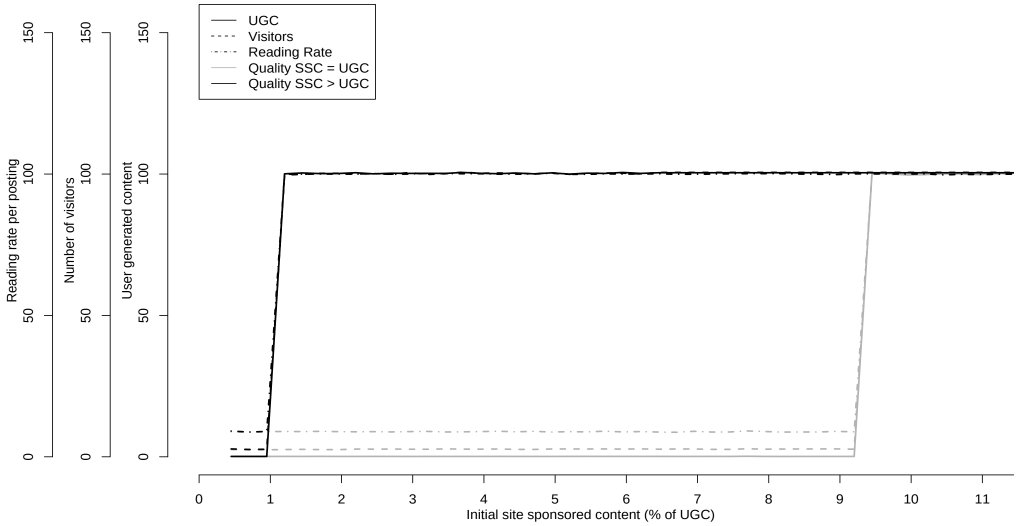
Next, we consider the role of SSC on tipping. That is, the site may invite sponsored content to "jump start" user activity. We consider two approaches to jump start the network, In the first, the site increases the initial SSC only. This strategy is analogous to sponsoring posts when the network is in the low-activity equilibrium and then stopping this practice after it tips. The impetus for this strategy is that the network will quickly become self-sustaining, such that the site no longer needs to bear the costs of sponsoring posts. Next, we contrast these results to a case wherein, instead of initial SSC only, the site continues sponsoring posts on a regular basis, and thus changes users' rational expectations regarding future SSC. We also consider the cases where the quality of SSC is either i) equal to the quality of UGC (by letting $\alpha_2^S/\alpha_2^U = 100\%$) or ii) substantially higher than that of UGC ($\alpha_2^S/\alpha_2^U = 10\%$). The results of this analysis are reported in Figure 5.

Results shown in Figure 5a suggest that the initial SSC equal to 9.3% of the UGC in the high-activity equilibrium is sufficient to tip the network when the quality of SSC is equal to that of UGC and that 1% is sufficient when the quality of SSC is much higher.

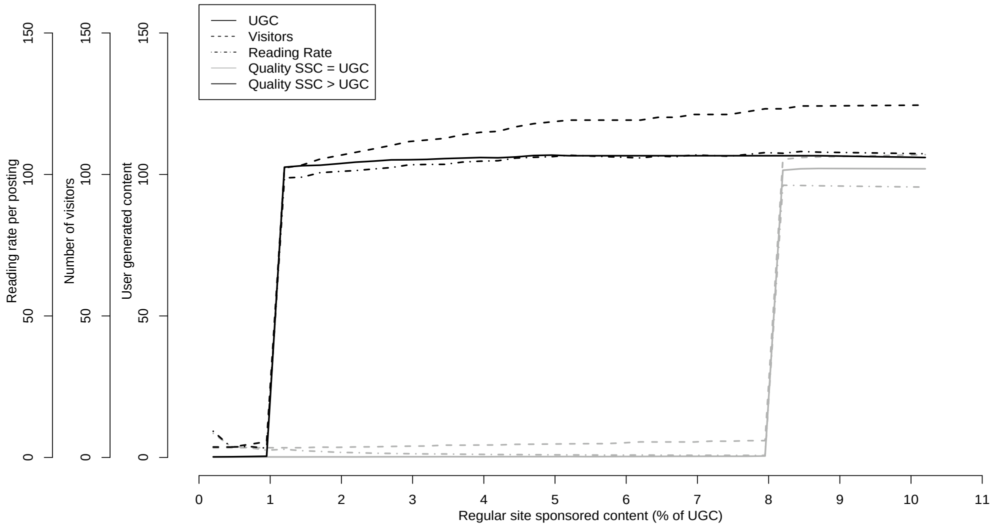
As evidenced in Figure 5b, the effect of posting sponsored content on regular basis tips the network at a level equivalent to 8% of the UGC in the high-activity equilibrium when site and user contents are equal in quality. The tipping point decreases further to 1% when the quality of SSC is much higher. The low tipping point arise because users perceive a regular stream of high quality content availability. Hence, we find it is possible for the site to tip the network with a relatively smaller number of high quality posts, so long as the site is committed to sponsoring content for an indefinite period of time.

Finally, we combined the two site strategies; initial and regular SSC to assess potential synergies in inducing the network to tip. Of interest is whether these approaches are complements or substitutes. If the former, it suggests that both should be used. If the latter, it suggests one or the other should be used. Figure 6 presents the results, which suggest that the two strategies are substitutes, and that there is no synergy. Thus, it appears that the two approaches are somewhat redundant and that the firm should pick one strategy or the other to tip the network, depending on the relative long-term costs of each.

To conclude in terms of jump starting the network, we contrast three strategies: i) enlist heavy users to post, perhaps through marketing via targeted advertising or incentives, ii) the site sponsoring initial posts only, and iii) the site taking a more regular route to sponsoring posts. The first option takes the most posts to tip the network, and the last strategy the least. The key reason it takes more user than site posts to tip the network pertains to the diminishing marginal returns to user posting. When users' initial stock increases, the marginal effect of their next posting decreases. With the site sponsors posts instead, the value of SSC remains high enough to attract readers, but the individual user stock is sufficiently low, so their incentive to post is greater. Because of this, the site tips more quickly.



(a) Effect of initial SSC strategy when initial UGC is set to zero.



(b) Effect of regular SSC strategy when initial UGC is set to zero.

Figure 5: Effect of sponsored content strategies. The horizontal axes represent initial and/or regular SSC as percentages of the UGC in the high-activity equilibrium. The vertical axis depicts the steady state site activity measures.

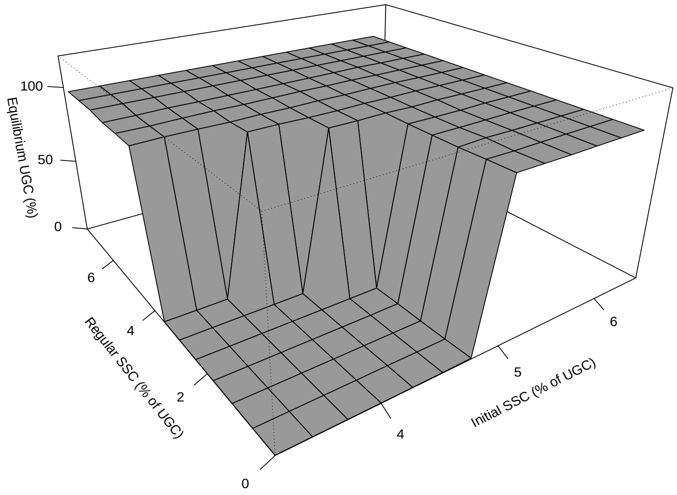
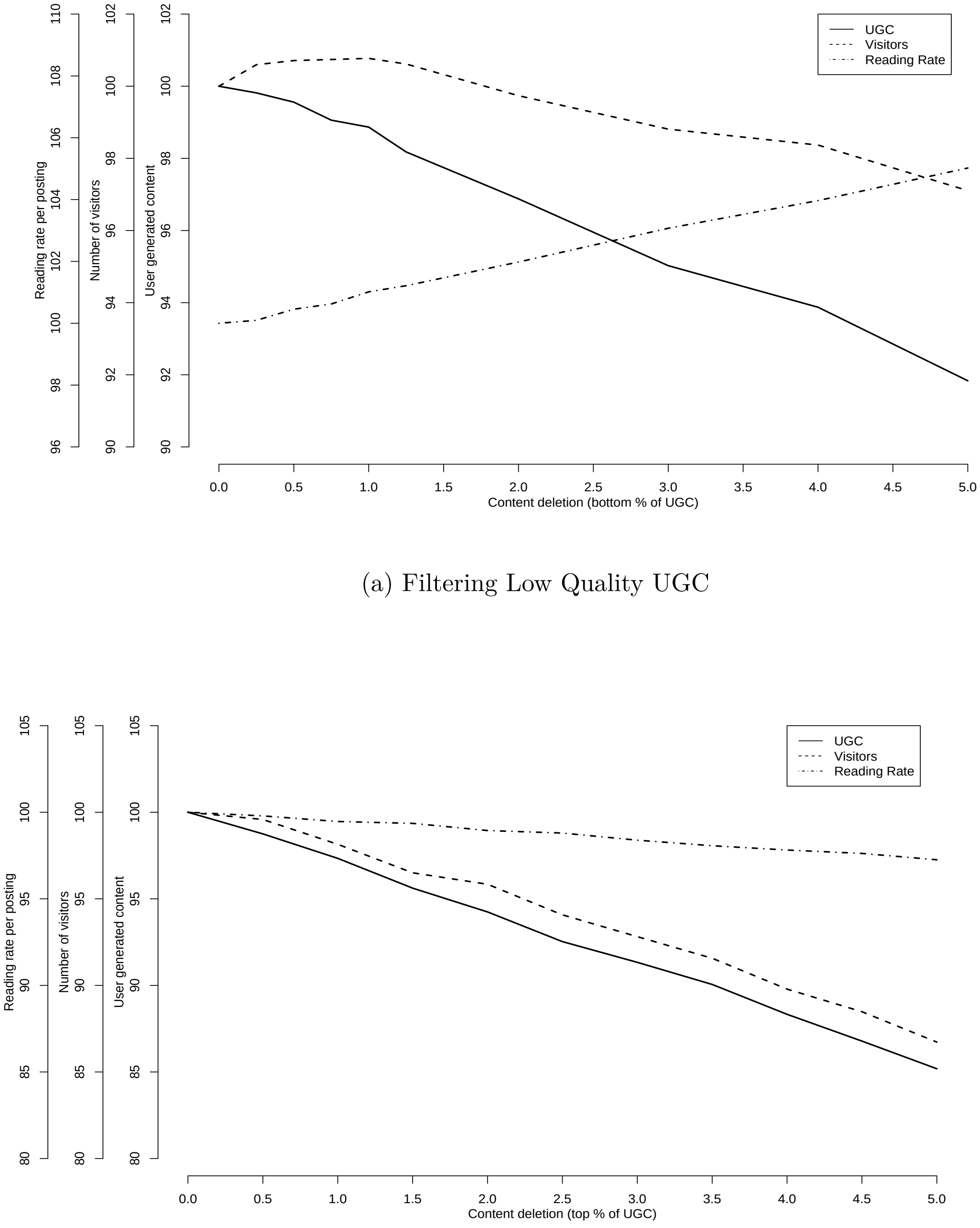


Figure 6: Interaction between initial and regular SSC for tipping when UGC is set to zero. The vertical axis depicts the steady state site participation rate as a percent of the UGC in the high-activity equilibrium.

The preferred strategy would be incumbent upon the relative cost of each. Contrasting whether the site should jump start with just initial posts (Figure 5a) or a constant stream of posts (Figure 5b), it is clear the latter is more effective at tipping. However, the cost of the latter strategy is borne each period rather than once. This fact would imply a tendency to favor a strategy of sponsored posts early for firms with low discount rates. When contrasting whether to use initial UGC or SSC, much depends on the relative cost of attracting each. For example, were the site able to sponsor high-quality posts at a reasonable cost in a short period of time, this would favor the strategy of such posts at the network's start, and then withdrawing completely after the network tips to the favorable equilibrium. Collectively, our counterfactual analyses indicate that site strategies have limited impact on the current steady state level of user engagement, but a more profound impact on ensuring the network takes off.



(a) Filtering Low Quality UGC

(b) Filtering High Quality UGC

Figure 7: Managing UGC Quality

7.3 Managing UGC Quality

Finally, we consider the possibility that a site filters user content, either by removing low quality posts or by removing high quality posts. The former case happens when sites remove offensive posts (including profanity or trolling). The latter case can occur when the site tightens enforcement of content that might violate fair use, as often happens on `Youtube.Com` when users pirate content. The results of this counterfactual are presented in Figure 7

Figure 7a considers the filtering of low quality UGC. Predictably, reading rates increase as average quality increases. However, knowing that content is subject to removal, users tend to generate less material on average. Stated differently, content remains as costly to produce, but it generates less expected utility. Given users create less content in response to having it filtered, the total drop in UGC will be greater than the amount of content removed. The two forces (better quality, but less content available and less utility from posting) trade off in terms of overall participation. In terms of improving overall traffic (number of visitors), it appears optimal to only filter a small amount of content, on the order of 1% to 1.25%.

The effect of filtering high quality UGC is quite asymmetric from that of filtering low quality UGC, as shown in Figure 7b. Unambiguously, deleting high quality content makes the site worse off in traffic, amount of available content and reading rate. However, the effects are slight if small amounts are deleted. Filtering the top 0.5% of content causes the number of the visitors to drop by only 0.4% and the reading rate to drop by 0.2%.

8 Conclusions

Recent advances in technology and media have enabled user generated content sites to become an increasingly prevalent source of information for consumers as well as a channel for advertisers to reach users of these sites. Hence, the factors driving the use of these networks is of a topical concern to marketers. In this paper, we consider how content, readership and site policy drive the evolution of content and readership on these sites.

Since our goal is to develop prescriptive and theoretical insights regarding user engagement on UGC platforms, we build upon the existing literature on social participation by developing a dynamic structural model to explore these effects. Individual reading behavior is developed from a model of information search that relates reading to the overall level of content on the site. Individual content generation is assumed to reflect the utility that users receive from the number of others reading the posts. Underpinning these two behaviors are users' beliefs regarding how the aggregate amount of content and readership on the platform evolve. These beliefs stem from the rational expectations equilibrium model whereby the evolution of aggregate reading and content states is assumed to be consistent with the aggregation of individual level reading and contribution decisions across the population.

Our paper makes several contributions. On a methodological front, we develop a dynamic structural model of UGC. Of future interest, this approach can be applied to assess the formation or dissolution of similar networks, such as academic journals (readers and authors), social media sites, blogs and so forth. Moreover, we extend the approximate aggregation approach along multiple dimensions, including i) enabling a single unit of supply to be consumed *concurrently* by many, ii) accommodating both *continuous* and discrete behaviors and iii) applying computational advances to enhance the *scale* of the problem solved. As a result, our approach facilitates the computation of a rational expectations equilibrium in the face of a large number of heterogeneous agents. Our advances could prove useful in other contexts in marketing and economics wherein firms face heterogeneous consumers.

On a theoretical dimension, we explore the tipping effects. We find that the potential exists for multiple equilibria depending upon whether initial usage can cross a sufficient threshold to attract participation. Another theoretical insight is that user and site sponsored content can serve as strategic complements or substitutes depending on whether the primary demand effect of content (attracting more users) dominates the secondary demand effect (splitting readers). An analogous argument can be constructed for past and current posts as their durability increases.

On a substantive domain, we consider a number of policy prescriptions to advise the site. First, we consider the role of sponsored content on user participation in a mature network. On the one hand, sponsored posts attract more readers, thereby growing the network. On the other hand, these posts are competitive with other users' posts. Overall, we conclude that the former effect predominates and the site can increase participation if sponsored posts are of sufficiently higher quality. Second, we consider the effect of sponsored and user content in jump starting a network. We find that the site can tip its network to a self-sustaining state by either incentivizing users to post or using sponsored content. We compare the strategy of continual posting SSC with the one wherein the site stops posting SSC once the network tips. Which strategy is most effective (initial user posts, initial sponsored posts, or regular sponsored posts) is incumbent upon the relative costs of each strategy. However, we note that a strategy of jump starting a network means a site can stop bearing the cost of sponsoring posts once the network tips. When sponsored post quality is sufficiently high, it only takes about 1% of the observed posting levels in the mature network to tip it. Perhaps this is one key reason that this latter approach was the one chosen by `Soulrider.com` to grow its network. Our study offers evidence of the efficacy of this strategy. Finally, we consider the impact of filtering user content, either by removing low quality or high quality posts. While filtering a small amount of very low quality content can increase site traffic, filtering a large amount of moderately poor quality content has an unambiguously deleterious effect on traffic. Likewise deleting high quality content makes the site worse off, although the effects are slight.

Several opportunities for extensions are present. First, the potential for competition exists for forum sites and extending our work to competing platforms would be of interest. Second, it would useful to extend our model to capture heterogeneity in content information in order to explore what information is most relevant in increasing site engagement. Related, certain leading content creators generate large followings and measuring the effect of lead users is of practical interest. Of note, these extensions potentially involve a considerable expansion of requirement for numerical computation in order to become feasible. Finally, our analysis considers a site where posts are not rated. The ratings of posts provide another incentive to post and would likely enter a joint posting-rating utility function. Owing to the prevalence of sites with rated content, this is an interesting future direction.

In sum, we hope that our research will lead to additional innovations in both user generated content and the application of the rational expectations equilibrium with approximate aggregation in marketing.

References

Albuquerque, Paulo, Polykarpos Pavlidis, Udi Chatow, Kay-Yut Chen, Zainab Jamal, Kok-Wei Koh, Andrew Fitzhugh. 2010. Evaluating promotional activities in an online two-sided market of user-generated content. *SSRN eLibrary* .

Ansari, Asim, Oded Koenigsberg, Florian Stahl. 2011. Modeling multiple relationships in social networks. *Journal of Marketing Research* **48**(4) 713 – 728.

Bughin, Jacques R. 2007. How companies can make the most of user generated content. *McKinsey Quarterly* 1–4.

Bulte, Christophe Van Den. 2007. *Social networks and marketing*. Marketing Science Institute, Cambridge MA.

Chevalier, Judith A., Dina Mayzlin. 2006. The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research* **43**(3) 345–354.

Chintagunta, Pradeep K., Ronald L. Goettler, Minki Kim. 2012. New drug diffusion when forward-looking physicians learn from patient feedback and detailing. *Journal of Marketing Research* **49**(6) 807–821.

Clarke, Darral G. 1976. Econometric measurement of the duration of advertising effect on sales. *Journal of Marketing Research* **13**(4) 345–357.

Dellarocas, Chrysanthos. 2006. Strategic manipulation of internet opinion forums: Implications for consumers and firms. *Management Science* **52**(10) 1577–1593.

Duan, W., B. Gu, A. B. Whinston. 2008. Do online reviews matter?: An empirical investigation of panel data. *Decision Support Systems* **45**(4) 1007–16.

Dubé, Jean-Pierre, Jeremy T. Fox, Che-Lin Su. 2012. Improving the numerical performance of static and dynamic aggregate discrete choice random coefficients demand estimation. *Econometrica* **80**(5) 2231–2267.

Dubé, Jean-Pierre, Günter J. Hitsch, Puneet Manchanda. 2005. An empirical model of advertising dynamics. *Quantitative Marketing and Economics* **3**(2) 107–144.

Dubé, Jean-Pierre H., Günter J. Hitsch, Pradeep K. Chintagunta. 2010. Tipping and concentration in markets with indirect network effects. *Marketing Science* **29**(2) 216–249.

Ghose, Anindya, Sang Pil Han. 2011. A dynamic structural model of user learning on the mobile Internet. *SSRN eLibrary* .

Ghosh, Arpita, Preston McAfee. 2011. Incentivizing high-quality user-generated content. *Proceedings of the 20th International Conference on World Wide Web*. ACM, 137–146.

Hartmann, Wesley R. 2010. Demand estimation with social interactions and the implications for targeted marketing. *Marketing Science* **29**(4) 585–601.

Hennig-Thurau, Thorsten, Kevin P. Gwinner, Gianfranco Walsh, Dwayne D. Gremler. 2004. Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the internet? *Journal of Interactive Marketing* **18**(1) 38–52.

Huang, Yan, Param V. Singh, Anindya Ghose. 2011. A structural model of employee behavioral dynamics in enterprise social media. *SSRN eLibrary* .

Iyengar, Raghuram, Christophe Van den Bulte, Thomas W. Valente. 2011. Opinion leadership and social contagion in new product diffusion. *Marketing Science* **30**(2) 195–212.

Karat, Clare-Marie, Christine Halverson, Daniel Horn, John Karat. 1999. Patterns of entry and correction in large vocabulary continuous speech recognition systems. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 568–575.

Katona, Zsolt, Peter Pal Zubcsek, Miklos Sarvary. 2011. Network effects and personal enfluences: The diffusion of an online social network. *Journal of Marketing Research* **48**(3) 425 – 443.

Katz, Michael L., Carl Shapiro. 1994. Systems competition and network effects. *Journal of Economic Perspectives* **8**(2) 93–115.

Katz, Michael L., Carl Shapiro. 1998. Antitrust in software markets. Jeffrey A. Eisenbach, Thomas M. Lenard, eds., *Competition, innovation and the Microsoft monopoly: Antitrust in the digital marketplace*. Kluwer Academic Publishers, 29–81.

Krusell, Per, Anthony A. Smith. 1998. Income and wealth heterogeneity in the macroeconomy. *The Journal of Political Economy* **106**(5) 867–896.

Lee, Donghoon, Kenneth I. Wolpin. 2006. Intersectoral labor mobility and the growth of the service sector. *Econometrica* **74**(1) 1–46.

Liebowitz, S. J., Stephen E. Margolis. 1994. Network externality: An uncommon tragedy. *The Journal of Economic Perspectives* **8**(2) 133–150.

Magnac, Thierry, David Thesmar. 2002. Identifying dynamic discrete decision processes. *Econometrica* **70**(2) 801–816.

Manski, Charles F. 1993. Dynamic choice in social settings: Learning from the experiences of others. *Journal of Econometrics* **58**(1) 121–136.

Mela, Carl F., Sunil Gupta, Donald R. Lehmann. 1997. The long-term impact of promotion and advertising on consumer brand choice. *Journal of Marketing Research* **34**(2) 248–261.

Moe, Wendy W., David A. Schweidel. 2012. Online product opinions: Incidence, evaluation, and evolution. *Marketing Science* **31**(3) 372 – 386.

Nair, Harikesh S, Puneet Manchanda, Tulikaa Bhatia. 2010. Asymmetric social interactions in physician prescription behavior: The role of opinion leaders. *Journal of Marketing Research* **47**(5) 883 – 895.

Nardi, Bonnie A., Diane J. Schiano, Michelle Gumbrecht, Luke Swartz. 2004. Why we blog. *Communications of the ACM* **47**(12) 41–46.

Nov, Oded. 2007. What motivates wikipedians? *Communications of the ACM* **50**(11) 60–64.

Preece, Jenny, Blair Nonnecke, Dorine Andrews. 2004. The top five reasons for lurking: Improving community experiences for everyone. *Computers in Human Behavior* **20**(2) 201–223.

Ransbotham, Sam, Gerald C. Kane, Nicholas H. Lurie. 2012. Network characteristics and the value of collaborative user-generated content. *Marketing Science* **31**(3) 387–405.

Rust, John. 1994. Structural estimation of Markov decision processes. Robert F. Engle, Daniel L. McFadden, eds., *Handbook of Econometrics*, vol. 4. North-Holland. Amsterdam., 3081–3143.

Ryan, Stephen P., Catherine Tucker. 2012. Heterogeneity and the dynamics of technology adoption. *Quantitative Marketing and Economics* **10**(1) 63–109.

Shriver, Scott K., Harikesh S. Nair, Reto Hofstetter. 2013. Social ties and user generated content: Evidence from an online social network. *Management Science* **59**(6) 1425–1443.

Stephen, Andrew T., Oliviet Toubia. 2010. Deriving value from social commerce networks. *Journal of Marketing Research* **47**(2) 215–228.

Stigler, George J. 1961. The economics of information. *The Journal of Political Economy* **69**(3) 213–225.

Su, Che Lin, Kenneth L. Judd. 2010. Structural estimation of discrete-choice games of incomplete information with multiple equilibria. *Proceedings of the Behavioral and Quantitative Game Theory: Conference on Future Directions*. BQGT '10, ACM, New York, NY, USA, 39:1–39:1.

Zhang, Kaifu, Theodoros Evgeniou, V. Padmanabhan, Emile Richard. 2012. Content contributor management and network effects in a ugc environment. *Marketing Science* **31**(3) 433–447.

Web Appendix

A Aggregate Reading Rate

Here we show that the expected amount of reading per post y_t defined in equation (10) can be closely approximated by the observed amount of reading per post. Given the expected amount of reading of a user, we obtain the aggregate expected amount of reading by all users,

$$R_t = E\left(\sum_{i=1}^{M} n_{it} r_{it}\right) = \sum_{i=1}^{M} E\left(n_{it} r_{it}\right), \tag{A1}$$

where M is the total number of users.[18] When we apply the latent segment model, the expected amount of reading of any user i is

$$\begin{aligned} E\left(n_{it} r_{it}\right) &= E\left[E\left(n_{it} r_{it} | n_{it}, \zeta_i\right)\right] \\ &= \int_{s_{it}} \sum_{j=1}^{J} p_j p\left(n_{it} = 1 | s_{it}\right) E\left(r_{it} | \bar{\zeta}_j, n_{it} = 1\right) dF\left(s_{it}\right), \end{aligned} \tag{A2}$$

where $F\left(s_{it}\right)$ is the stationary distribution of the state variables s_{it}, and $p(n_{it} = 1|s_{it})$ is the probability that the user i visits the site at period t defined in Section C.3. By substituting A2 into A1, we have expected aggregate amount of reading as

$$R_t = M \int_{s_{it}} \sum_{j=1}^{J} p_j p\left(n_{it} = 1 | s_{it}\right) \frac{\alpha_1 + \bar{\zeta}_j - \bar{\kappa}_{1j} w_t}{\alpha_2 / K_t + \bar{\kappa}_{2j}} dF\left(s_{it}\right).$$

The observed total amount of reading, denoted by $\tilde{R}_t$, is defined as and it follows that

$$E\left(\tilde{R}_t\right) = E\left(E\left(\tilde{R}_t | n_{it}, \zeta_i\right)\right) = M \int_{s_{it}} \sum_{j=1}^{J} p_j p\left(n_{it} = 1 | s_{it}\right) \frac{\alpha_1 + \bar{\zeta}_j - \bar{\kappa}_{1j} w_t}{\alpha_2 / K_t + \bar{\kappa}_{2j}} = R_t.$$

When the number M is large, $\tilde{R}_t / M$ is approximately equal to $E\left(n_{it} r_{it}^*\right) = R_t / M$ because of the law of large numbers. So the expected average amount of reading per post becomes

$$y_t = \frac{R_t}{K_t} = \frac{R_t / M}{K_t / M} \approx \frac{\tilde{R}_t / M}{K_t / M} = \frac{\tilde{R}_t}{K_t},$$

which implies that we can use the observed average amount of reading per post to approximate the expected one in our model when the number of users is very large.

[18] The number of users increased by 0.37% over the sample period of two months, which translates into a 2.2% rate of annualized growth rate. Despite this modest growth rate, we treat the market size, M, as fixed over time in our model, because some registered users may also drop from the site. The assumption of fixed market size is further justified, if the reading and posting behavior of regular users stays stationary over time. Our empirical analysis on the temporal movement of K_t indeed supports this stationary assumption.

B Rational Expectations Equilibrium Simulation

The following steps outline our approach to computing a rational expectations equilibrium for the policy simulations and theoretical analysis.

1. Set structural parameters for utilities and costs of site participation, reading, and writing. Impose bounds on state space of K_t, $\{k_{i,t}\}_{i=1}^{M}$, and y_t. Select grid points in the state space.

2. Guess the values for ω_0^K, ω_1^K, ω_2^K and $\omega_0^y, \omega_1^y, \omega_2^y$ in equations (23) in Section 4.6 and (24).

3. Solve for $p(n_{it} = 1|s_{it})$, $p(r_{it}|s_{it}, n_{it} = 1)$, and $p(a_{it}|s_{it}, n_{it} = 1)$. The solution to dynamic programming requires the value of y_t consistent with both aggregate reading and writing decisions (R_t and K_t). To get this value, we use the following steps:

   (a) Choose an arbitrary y_t^{old} and K_t^{old}

   (b) Compute $p(n_{it} = 1|s_{it})$, $p(r_{it}|s_{it}, n_{it} = 1)$, and $p(a_{it}|s_{it}, n_{it} = 1)$ and solve for decisions by users, $\{n_{it}, r_{it}, a_{it}\}_{i=1}^{N}$.

      i. Given y_t^{old}, we can solve for $p(a_{it}|s_{it}, n_{it} = 1)$. We use Rust (1987) to solve $\tilde{EV}_i\left(s_{it}, a_{it}\right)$ and Chebyshev approximation to interpolate the expected value functions.

      ii. Given K_t^{old}, we can solve for individual-level optimal reading r_{it}^*.

      iii. Given $p(a_{it}|s_{it}, n_{it} = 1)$ and r_{it}^*, we can solve for $p(n_{it} = 1|s_{it})$.

   (c) Compute y_t^{new} and K_t^{new}. Check if $y_t^{old} = y_t^{new}$ and $K_t^{old} = K_t^{new}$. If the conditions hold then stop. If not, set $y_t^{old} = y_t^{new}$ and $K_t^{old} = K_t^{new}$ and iterate steps 3a-3c until convergence.

   (d) Solve for rational expectations by using OLS estimation for

$$\begin{aligned} K_t &= \tilde{\omega}_0^K + \tilde{\omega}_1^K K_{t-1} + \tilde{\omega}_2^K w_t + \varepsilon_t^K, \\ y_t &= \tilde{\omega}_{0t}^y + \tilde{\omega}_1^y K_t + \tilde{\omega}_2^y w_t. \end{aligned}$$

4. Check if $\omega_0^K, \omega_1^K, \omega_2^K$ and $\omega_0^y, \omega_1^y, \omega_2^y$ are close to $\tilde{\omega}_0^K, \tilde{\omega}_1^K, \tilde{\omega}_2^K$ and $\tilde{\omega}_0^y, \tilde{\omega}_1^y, \tilde{\omega}_2^y$. If the conditions hold then stop. If not, replace $\omega_0^K, \omega_1^K, \omega_2^K$ and $\omega_0^y, \omega_1^y, \omega_2^y$ with $\tilde{\omega}_0^K, \tilde{\omega}_1^K, \tilde{\omega}_2^K$ and $\tilde{\omega}_0^y, \tilde{\omega}_1^y, \tilde{\omega}_2^y$. Iterate steps 2-3 until convergence.

Note that in estimation, the aggregate state transitions are observed and assumed to reflect rational expectations in the current equilibrium, so no iteration is necessary to achieve the rational expectations equilibrium. In policy simulations and theoretical analysis, however, we need to iterate to obtain it.

C Model Estimation

C.1 Estimating Content Consumption Model

We assume that there are J segments and if user i is in the j-th segment, we have the reading model

$$r_{it} = \frac{\alpha_1 - \bar{\zeta}_j - \bar{\kappa}_{1j} w_t}{\alpha_2 / K_t + \bar{\kappa}_{2j}} \nu_{it} \tag{A3}$$

If we assume that ν_{it} has the exponential distribution, the likelihood function for r_{it} given i in segment j is

$$\text{Exponential}\left(r_{it} | \frac{\alpha_1 - \bar{\zeta}_j - \bar{\kappa}_{1j} w_t}{\alpha_2 / K_t + \bar{\kappa}_{2j}} \right)$$

If we do not know segment membership of i, the likelihood becomes the following finite mixture distribution

$$\sum_{j=1}^{J} p_j \text{Exponential}\left(r_{it} | \frac{\alpha_1 - \bar{\zeta}_j - \bar{\kappa}_{1j} w_t}{\alpha_2 / K_t + \bar{\kappa}_{2j}} \right).$$

C.2 Estimating Content Generation Model

The value function for the posting decision in the form of Bellman's equation is

$$V_i(s_{it}, \varepsilon_{it}) = \max_{a_{it} \in A} \{ u^P(a_{it}) - \bar{c}_{it}^P(a_{it}) + \varepsilon_{it}(a_{it}) + \beta E[V_i(s_{i,t+1}, \varepsilon_{i,t+1}) | s_{it}, a_{it}] \}.$$

where $E[V_i(s_{i,t+1}, \varepsilon_{i,t+1}) | s_{it}, a_{it}]$ represents the expected future value given the current states, and $u^P(a_{it}) - \bar{c}_{it}^P(a_{it}) + \varepsilon_{it}(a_{it})$ represents the current period net utility of action a_{it}. To derive the probability of observing a user posting a specific number of posts under this optimal decision rule, we first use the conditional independence assumption for ε_{it} and s_{it} to find

$$\tilde{EV}_i(s_{it}, a_{it}) =$$
$$\int_{s_{i,t+1}} \log \left\{ \sum_{a_{i,t+1} \in A} \exp \left[u^P(a_{i,t+1}) - \bar{c}^P_{i,t+1}(a_{i,t+1}) + \beta \tilde{EV}_i \left(s_{i,t+1}, a_{i,t+1} \right) \right] \right\} p(s_{i,t+1}|s_{it}, a_{it}) ds_{i,t+1}.$$

One can solve this fixed point equation to obtain solutions for $\tilde{EV}_i(s_{it}, a_{it})$ (Rust 1987, 1994). These solutions are then used in the computation of posting and participation probabilities. Specifically, the optimal posting decision is to choose a_{it} if and only if

$$\begin{aligned} & u^P(a_{it}) - \bar{c}^P_{it}(a_{it}) + \varepsilon_{it}(a_{it}) + \beta \tilde{EV}_i(s_{it}, a_{it}) \geq \\ & \qquad u^P(a'_{it}) - \bar{c}^P_{it}(a'_{it}) + \varepsilon_{it}(a'_{it}) + \beta \tilde{EV}_i(s_{it}, a'_{it}), \forall a'_{it} \neq a_{it} \in A, \end{aligned} \tag{A4}$$

by which we derive the probability of writing a_{it} content postings conditional on site participation as

$$P(a_{it}|s_{it}, n_{it} = 1) = \frac{\exp(u^P(a_{it}) - \bar{c}^P_{it}(a_{it}) + \beta \tilde{EV}_i(s_{it}, a_{it}))}{\sum_{a'_{it} \in A} \exp(u^P(a'_{it}) - \bar{c}^P_{it}(a'_{it}) + \beta \tilde{EV}_i(s_{it}, a'_{it}))}. \tag{A5}$$

One key component of estimation is to approximate the expected value functions in equation (16). This task is nontrivial for our model, because our state variables are mostly continuous with a wide support. Moreover, the control variable can take high-order discrete values. For this reason, we use Chebyshev approximation to approximate the expected value functions as described in (Dubé et al. 2012, Miranda and Fackler 2002). Chebyshev approximation uses polynomial interpolation to approximate the expected value functions:

$$\tilde{EV}_j(s, a) \approx \psi \Gamma(s, a).$$

We can then rewrite the Bellman equation in the fixed point algorithm as a function of the interpolated functions

$$\psi \Gamma(s, a) = \int_{s'} \log \left(\sum_{a' \in A} \exp \left\{ u^P(a_{it}) - \bar{c}^P_{it}(a_{it}) + \beta \psi \Gamma \left(s', a' \right) \right\} \right) \cdot p(s'|s, a) ds'.$$

To compute the right-hand side of the above equation, we need to numerically evaluate an indefinite integral with respect to state transition probabilities of aggregate stock of posting. Since we use a normal distribution to model these probabilities, the Gauss-Hermite quadrature can be used to approximate the integration in the Bellman equation above. The Gauss-Hermite quadrature allows us to evaluate the integrand at fewer points than, for example, a Monte Carlo integration.

Once we compute both sides of the fixed point equation, we can formulate the following constraints to be used for our estimation based on the MPEC approach (Su and Judd, 2010):

$$R(s, a; \psi) = \psi\Gamma(s, a) - \int \log\left(\sum_{a' \in A} \exp\left\{u^P(a_{it}) - \bar{c}^P_{it}(a_{it}) + \beta\psi\Gamma\left(s', a'\right)\right\}\right) \cdot p(s'|s, a)ds' = 0.$$

By approximating the expected value functions, we can transform a dynamic discrete choice model into a static computational equivalent and use a maximum likelihood estimation to recover the structural parameters of our interest.

The joint likelihood of reading and posting for all individuals is then[19]

$$\left\{\prod_{i=1}^{M}\sum_{j=1}^{J} p_j \prod_{t=1}^{T} \text{Exponential}\left(r_{it}|\frac{\alpha_1 - \bar{\kappa}_{1j}w_t - \bar{\zeta}_j}{\alpha_2/K_t + \bar{\kappa}_{2j}}\right) \frac{\exp\left(u^P(a_{it}) - \bar{c}^P_{it}(a_{it}) + \tilde{EV}_j(s_{it}, a_{it})\right)}{\sum_{a'_{it} \in A} \exp\left(u^P(a_{it}) - \bar{c}^P_{it}(a_{it}) + \tilde{EV}_j(s_{it}, a'_{it})\right)}\right\}. \tag{A6}$$

The direct MLE approach is applied to estimate the parameters. To compute the standard errors of parameter estimates in the posting model, we use nonparametric bootstrapping. Note that we allow for heterogeneity for reading and posting costs using finite mixture models, which makes it difficult to implement nonparametric bootstrapping for computing standard errors due to the label switching problem. Geweke and Keane (1997) propose labeling restrictions that prevent the components of the mixture from interchanging across bootstrapped samples. For example, segments can be ordered according to their sizes to preserve segment labels consistently across bootstrapped samples.

C.3 Estimating Site participation Model

Lastly, we have the likelihood function for site-participation in equation (22) as

$$\left\{\prod_{i=1}^{M}\sum_{j=1}^{J} p_j \prod_{t=1}^{T} \left[P(n_{it} = 1|s_{it})^{n_{it}} P(n_{it} = 0|s_{it})^{1-n_{it}}\right]\right\}, \tag{A7}$$

which is also estimated by MLE.

[19] Note that the reading and posting model are jointly estimated, and these components are linked by the indirect effect of posting on reading and reading on posting.

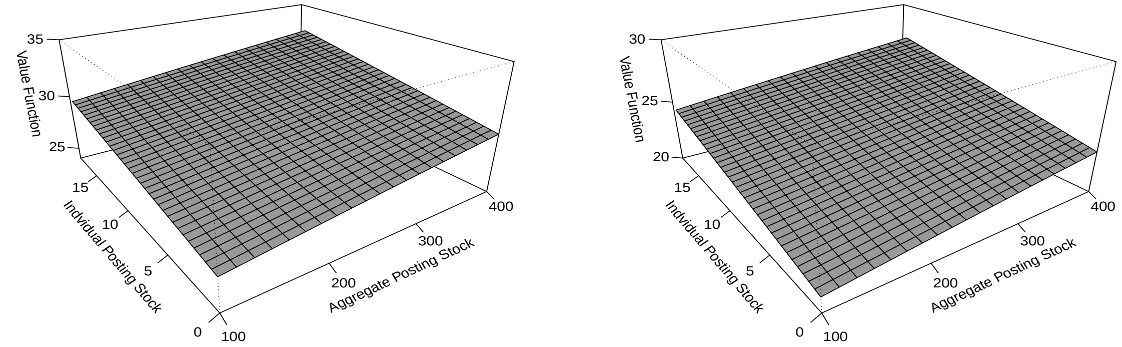


Figure 8: Estimated Value Function for Two Segments of Users

C.4 Estimated Value Functions

The post levels are treated as ordered numbers $\{0, 1, 2, 3, ..., A\}$ in the posting utility and mean cost functions. Hence, the posting utility and mean costs are actually ordered. To demonstrate the estimated value functions are monotonic, we depict the estimated value functions plotted against its two ordered state variables, aggregate UGC and individual post stock below.

D Theoretical Implications by Simulation

In this section, we explore some of the theoretical properties of our model. Specifically, we assess i) convergence to the defined rational expectations equilibrium in Section 4.6 and ii) how the model's parameters and aggregate states influence the network's user content and readings in equilibrium. This analysis considers the role of initial content on network size, the effects of reading and posting costs and content stock decay on content generation, and the self-fulfilling prophecies under rational expectations.

D.1 Simulation Design

We simulate 2 segments of 3000 and 6000. We let both segments have the same cost of reading and heterogeneous costs ($\bar{\xi_j}$) of content generation. To simply the simulation, we assume that there is no cyclical effect ($\bar{\tau}_j = 0$ and $\bar{\kappa}_{1j} = 0$). The reading cost parameters $\alpha_1 - \kappa_1 = 0.1$, $\alpha_2 = 1$ and $\kappa_{2j} = 0.0015$ imply that a post stock of $K_t^U = 10,000$ will induce an individual user to read 62.5 different postings per period. We let the cost of posting for Segment 1 be $\bar{\xi_1} = 0.1$ and Segment 2 be $\bar{\xi_2} = 5$. Note that $\bar{\xi_2}$ is 50 times of $\bar{\xi_1}$, which implies that Segment 2 has a much higher cost of posting, and hence users in Segment 2 are likely to post much less than those in Segment 1. Indeed, we find in equilibrium a user in Segment 2 writes only about 2 postings in 100 periods whereas a user in Segment 1 writes about 350 posting in the same periods on average. We set the posting utility parameter $\gamma = 0.5$.

We endow every individual user with a randomly selected initial stock of user generated content. The initial aggregate stock of UGC is the summation of individual stocks plus a fixed initial stock of sponsored content. The discount parameter β in the utility of posting is set to be 0.98. The site sponsored content K_t^S is assumed to be exogenously set at 2000.

We simulate individual postings and amount of reading for 100 periods. We then use the aggregate number of postings to re-estimate the dynamic law of motion for the post stock, which will in turn lead to new value functions for both segments of user. The new value functions are used to simulate individual posting data again. This process is iterated until the law of motion for the posting converges. From numerous repeated experiments, we found it takes fewer than 20 iterations to converge to the rational expectations equilibrium. For illustration purpose, we show an example where the decay parameter ρ is set to be 0.6.

D.2 Simulation Results

D.2.1 Initial Individual Stock

We select two different sets of values for the initial endowment of individual post stocks. The first set of values has the post stock equal to 3 for any individual in Segment 1 and 0.1 for Segment 2; the second has 8 for Segment 1 and 0.1 for Segment 2. Neither of these two sets of initial values are considered extremely high or low, so we expect they converge to the same equilibrium.

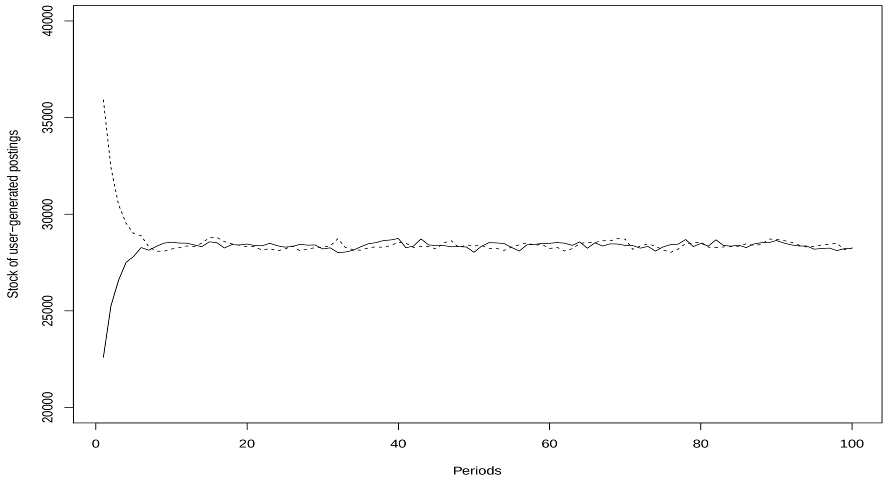


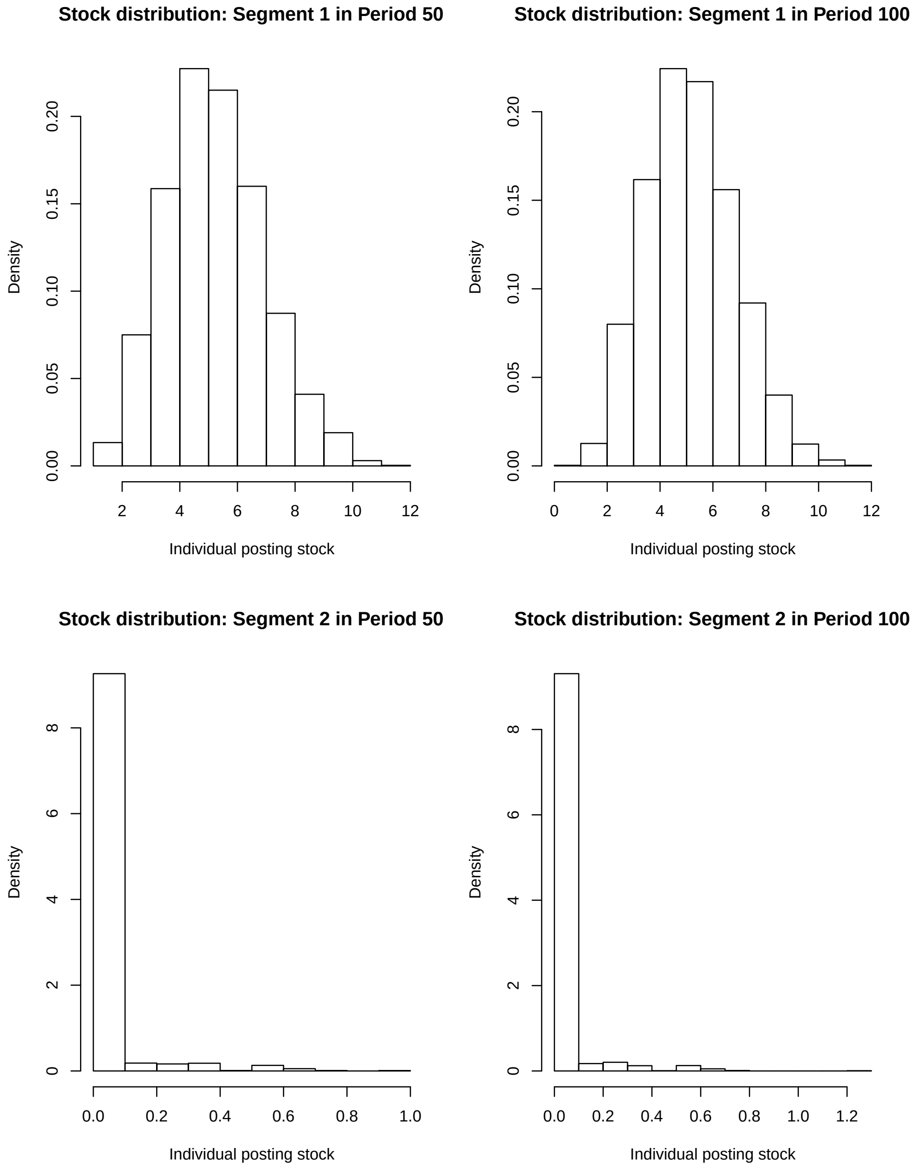
Figure 9: Convergence of aggregate UGC stock to the steady state from 2 different starting values.

In Figure 9, we plot the equilibrium path of the aggregate user generated postings (UGC) after the rational expectations equilibrium is achieved. We can see that the first set of initial values (solid curve) and the second (dashed curve) converge to the same steady-state aggregate UGC with small, random variations. We also find the same equilibrium parameter values in the equations(23) and (24) in Section 4.6. The UGC reaches the steady state after only about 10 periods.

Based on the theoretical model in Section 4.2, we expect that not only the aggregate UGC converges (shown in Figure 9), but also the distribution of individual post stocks would be constant in the steady state as well. Figure 10 shows the distribution (histogram) of individual post stocks of the two segments of site users in period 50 and 100, when the UGC has already reached the steady state. These histogram plots confirm our conjecture that these distributions are indeed invariant over time.

D.2.2 Degenerate Equilibrium

A potential equilibrium of our model is that all postings, reading and participation are zero. That is, the network will never expand unless some shock or intervention enables the network to tip from a non-zero state. For example, an extremely low user post stock can cause the


Stock distribution: Segment 1 in Period 50

Stock distribution: Segment 1 in Period 100

Stock distribution: Segment 2 in Period 50

Stock distribution: Segment 2 in Period 100

Figure 10: Distributions of individual user's post stocks of the two segments defined in Section D in steady state.

low reading and visiting rate, which can in turn cause even lower posting activity. In order to test this conjecture, we select a set of very low initial values for post stocks: 0.1 for both Segments 1 and 2. The dashed curve in Figure 11 demonstrates the result of this simulation which converges to the trivial equilibrium.

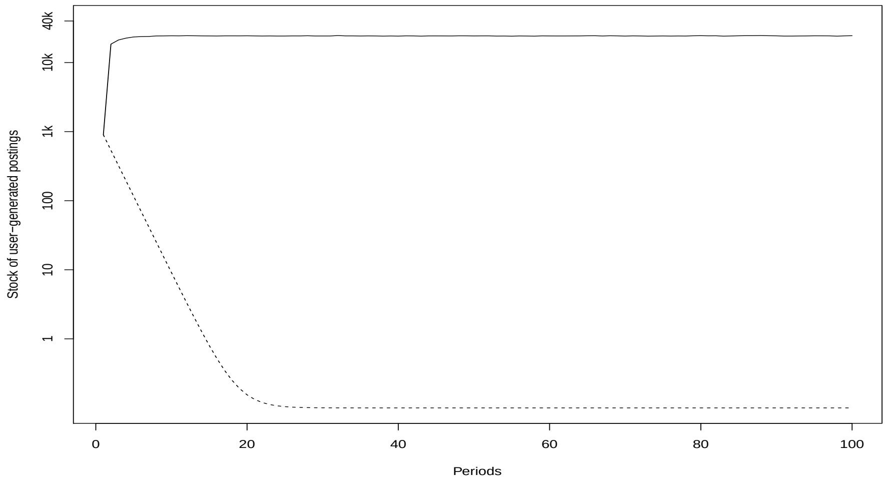


Figure 11: Convergence of the aggregate UGC stock to two different steady states from a common starting value when the initial individual post stock is 0.1 and (i) the site sponsored content (K_t^S) has the means equal to $20,000$ (solid curve) and (ii) $2,000$ (dashed curve).

D.2.3 Decay Parameter and Average Number of Postings per Person

The decay parameter ρ of site's postings implies two opposite effects on user posting activity. First, a lower decay rate (higher ρ) means that a post is more likely to be seen in the future, so a user has the incentive to post more. This also raises content available for readers thereby increasing site participation. However, higher ρ makes posting more "durable" and hence increases the aggregate post stock and decreases the rate of reading per post, which could cause a user to post less. The net effect of ρ is not clear directly from the utility function, because a closed-form derivative of the utility with respect to the decay parameter cannot be easily derived. Therefore, we discretize the space of the decay parameter ($\rho \in [0, 1]$) to ten equally spaced grid points (0.1, 0.2, ..., 0.9) and simulate the content and reading given these values.

In Figure 12, we depict the relationship between the decay parameter and the average number of postings per period per user in Segment 1 (solid curve) based on the simulation results. Figure 12 also shows the relationship between the decay parameter and the average reading per post y_t (dashed curve). From Figure 12, a higher decay parameter will *ceteris paribus* lower average reading per post thanks to the competitive effect of more durable postings. However, the average number of postings per user increases when the decay parameters ρ increases from 0.1 to 0.4 and decreases when ρ is above 0.5. This result is due to the two opposite effects of ρ on user activity.

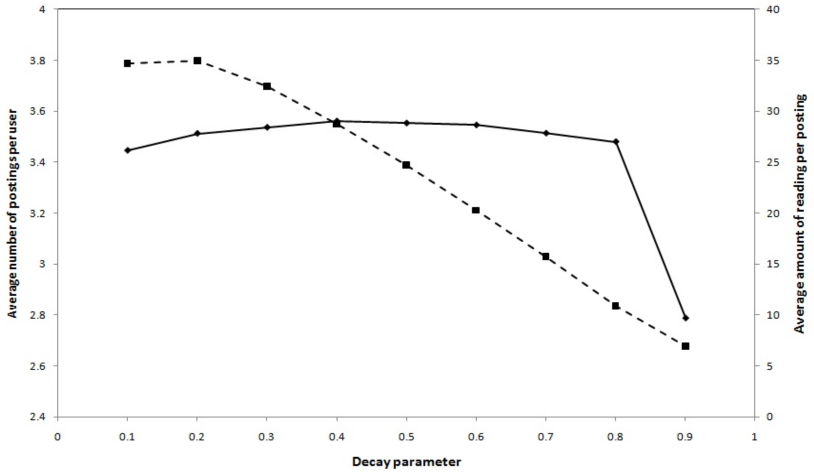


Figure 12: Average number of postings by individual users in steady state vs. the decay parameter ρ (solid curve) and the average reading per post y_t vs. the decay parameter ρ (dashed curve).

D.2.4 Self-fulfilling Prophecies

Owing to the formation of expectations regarding the aggregate state transitions in Equations (24) and (25), content generation and reading decisions are incumbent upon future beliefs. Of interest is the possibility that these beliefs become self-reinforcing. This issue can be explored by shocking these beliefs in the short-term (by varying the initial states and variances in the state transition equations) and the long-term (by varying the regression coefficients in the state transition equations) seeing how the evolution of content generation changes.

In order to test whether shocking short-term beliefs can lead to different long-term behaviors, i.e., converging to different equilibria of the model, we reset the initial belief about the aggregate UGC stock to 5%, 25%, 50%, 150%, and 200% of the actual stock and simulate the rational expectations equilibrium following the algorithm in Section 4.6. We find that all these simulations converge to the same equilibrium which has the same levels of mean UGC and number of visitors as in the observed data. We also reset the initial belief about the variance in the state transition equation for the aggregate UGC to 25%, 50%, 150%, 200% and 300% of the value estimated from the real data. All these simulations again converge to the original equilibrium. Hence, we conclude that shocking short term beliefs will not lead to self-fulfilling behavior.

To evaluate whether erroneous long-term beliefs about the transition rule of aggregate UGC can lead to different equilibrium, we set the initial value of the coefficient ω_1^K in equation 25 to 0.1, 0.2,...,0.9 and simulate their corresponding equilibria. We find that they converge to the same equilibrium in which the auto-regressive coefficient ω_1^K is 0.84. Therefore, erroneous long-term beliefs about the transition rule will not lead to self-fulfilling behavior. Note that the rational expectations equilibrium in our model is similar to that by Krusell and Smith (1998), who also found the absence of self-fulfilling behavior in their model.

D.2.5 Network Effect of Aggregate UGC

Thanks to its structural foundation and inclusion of direct and indirect network effects, our model evidences flexibility in characterizing aggregate UGC behaviors. We exemplify this point below by considering i) the marginal effect of additional content and ii) the role of initial stock. We explore other theoretical implications of our model in Appendix D.[20]

A key consideration in assessing the role of UGC on overall site traffic is the marginal effect of additional content on consumption. The indirect network effect of the aggregate UGC on an individual user's posting action is affected by the likelihood her post is read; that is, the numerator (aggregate reading) and denominator (aggregate postings) in equation (10). The numerator implies a greater likelihood of reading because more content, K_t, enhances the consumption experience. The denominator implies a competitive effect of K_t as content increasingly competes for users. As we show below, this indirect effect can be either positive or negative.

In Figure 13, we show two simulation examples, one where y_t is decreasing in K_t and another where it is decreasing. The decay rate ρ is 0.6 for the first example and 0.1 for the

[20] The appendix details i) convergence to the defined rational expectations equilibrium in Section 4.6 and ii) how the model's parameters influence the network's user content and readings in equilibrium, including the role of initial content on network size and the effect of content stock decay on content generation.

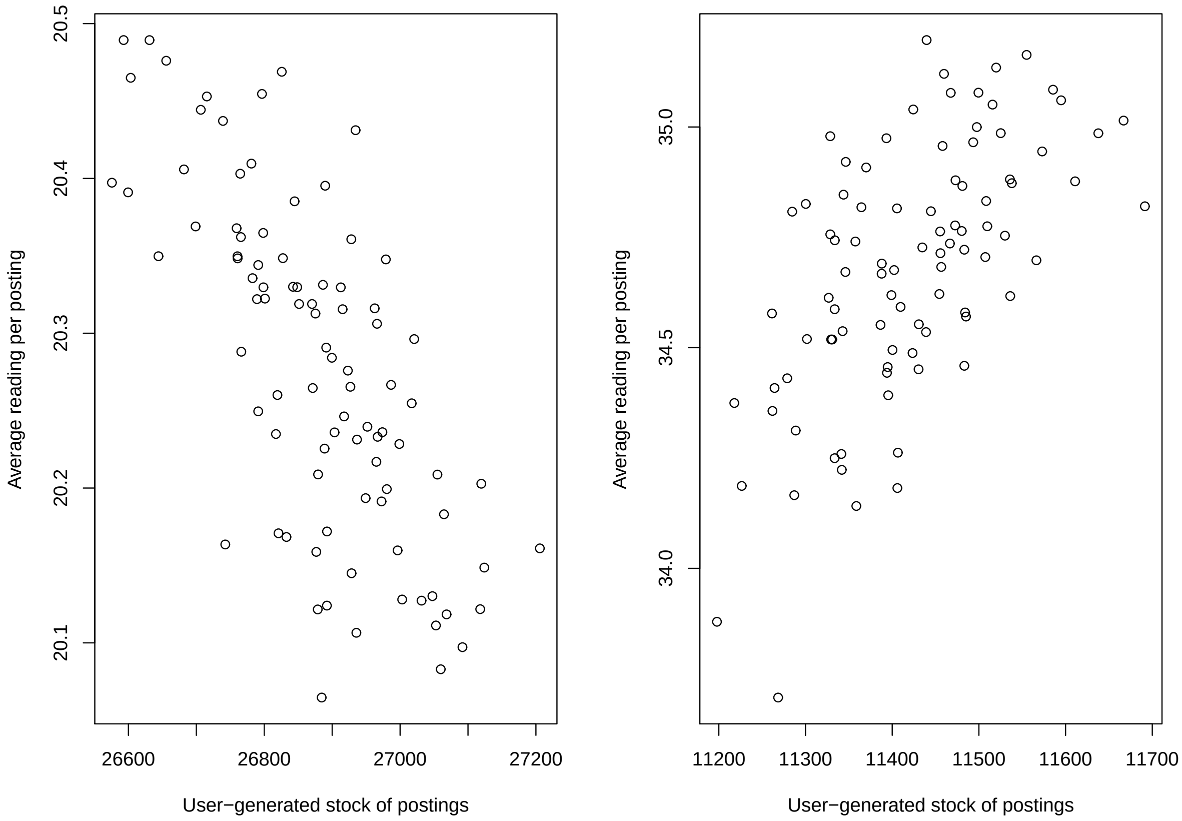


Figure 13: The relationship between average reading per post and aggregate UGC stock in equilibrium.

second: all the remaining parameter values are identical in the two examples.[21] We also find the relationship between y_t and K_t can switch sign if we adjust the ratio of population sizes of the two segments. Because the numerator is not a closed-form function of K_t, it is not clear under what conditions the network effect of K_t is positive. We conjecture that the positive indirect effect is more likely when there is a strong primary effect on site participation and that the negative indirect effect is more likely when the participation is already high.

Of note, a descriptive model of network effects (common in prior research) would lead to vastly different interpretations of the indirect network effect under these two scenarios. In

[21] This simulation considers two segments in our 100 period simulation, each of whom have the same cost of reading but vary in their posting costs and size (Segment 1 is smaller and has lower posting costs, consistent with the notion that a small number of users predominate the number of posts). Appendix D.1 details the specific parameters of our simulation.

contrast, the model we develop is sufficiently flexible to accommodate the change in the signs of these marginal effects, depending on the state of the network. The sign of the network effect at the current state of the network is especially important for the forum managing firms when they consider whether using sponsored content can increase traffic and UGC on their web sites.

D.3 Evidence of Dynamic Behavior

In this section, we note that the empirical patterns reported in Section 3.3.3 could be consistent with the theoretical implications of the model implied in Section 4.7. Specifically, we consider the effect of future costs on current content, future reading rates on current content, and stock on future content - holding all else constant.[22]

- Future Costs. First, a regression of the form $a_t = (\cdot) + \theta_\tau \tau_{t+1} + e_t$ implies that $\theta_\tau > 0$ is consistent with the Euler constraint. That is, as future costs increase, users accelerate content generation. Though we do not directly observe τ_t, content generation costs appear to be higher on the weekend (due to higher opportunity costs and perhaps more limited online access). Thus, if content generation (a_t) is higher (lower) the day before the weekend (week), this is consistent with strategic management of posts. Consistent with this logic, the effect of future weekend on current posting in Table 4 is negative (p-value < 0.01).

- Future Reading Rates. Content generation in the current should decrease when future reading rates increase as users delay posts. Hence, in a regression of $a_t = (\cdot) + \theta_y y_{t+1} + e_t$, we would expect to find that $\theta_y < 0$. As expected, we find the effect of future reading rates on content in Table 4 is negative (p-value < 0.05). .

- Individual Stock. Though it is not specifically an analysis of strategic behavior, higher stock in the current period leads to lower posts in the next period. This leads to a regression of the form $a_t = (\cdot) + \theta_k k_{t-1} + e_t$ where our theory implies $\theta_k < 0$. Consistent

[22] These regressions presume all else equal, but in fact the data series move jointly. For example, if future reading rates increase, entering stock might also vary. In light of this, we interpret our results somewhat conservatively.

with this conjecture, Table 4 indicates a strong negative correlation between lagged individual stock and content generation (p-value < 0.01).

In sum, these regressions might offer tentative support for the dynamic behavior observed in the model.

E The Effect of Site Sponsored Content on Reading

We define $E(Q^S_{[k_s]}|K^S_t, \Theta^S)$ as the expected quality of the k_sth highest quality sponsored post given a total of K^S_t sponsored posts and the quality distribution parameters $\Theta^S = \{U^S, L^S\}$. Likewise, we define $E(Q^U_{[k_u]}|K^U_t, \Theta^U)$ as the expected quality of the k_uth highest user post given a total of K^U_t user posts and the quality distribution parameters $\Theta^U = \{U^U, L^U\}$. We seek to compute, for any given combination of K^S_t and K^U_t, how a reader's utility varies with the level of site and user content. The combined utility from reading r^S_{it} sponsored posts and r^U_{it} user generated posts of the highest overall quality is given by

$$
\begin{aligned}
u^R(r^U_{it}, r^S_{it}) &= E\left(\sum_{k_s=K^S_t-r^S_{it}+1}^{K^S_t} Q_{[k_s]} + \sum_{k_u=K^U_t-r^U_{it}+1}^{K^U_t} Q_{[k_u]} | K^S_t, \Theta^S, K^U_t, \Theta^U \right) \\
&= \left(\left(U^S - L^S \right) \left\{ \frac{K^S_t + 1/2}{K^S_t + 1} r^S_{it} - \frac{1}{K^S_t + 1} \frac{r^{S^2}_{it}}{2} \right\} + L^S r^S_{it} \right) \\
&\quad + \left(\left(U^U - L^U \right) \left\{ \frac{K^U_t + 1/2}{K^U_t + 1} r^U_{it} - \frac{1}{K^U_t + 1} \frac{r^{U^2}_{it}}{2} \right\} + L^S r^S_{it} \right).
\end{aligned} \tag{A8}
$$

For any given $K^S_t$and K^U_t sufficiently large, this expression reduces to

$$
u^R(r^U_{it}, r^S_{it}) = \alpha^S_1 r^S_{it} - \frac{\alpha^S_2}{2K^S_t} r^{S^2}_{it} + \alpha^U_1 r^U_{it} - \frac{\alpha^U_2}{2K^U_t} r^{U^2}_{it}, \tag{A9}
$$

where $\alpha^S_1 = U^S$, $\alpha^S_2 = U^S - L^S$, $\alpha^U_1 = U^U$ and $\alpha^U_2 = U^U - L^U$. Under the assumption of quadratic reading costs, the total cost of reading is given by $\kappa_{1it} \left(r^S_{it} + r^U_{it} \right) + 1/2\kappa_{2i} \left(r^S_{it} + r^U_{it} \right)^2$. Adding the user heterogeneity in the reading utility, the optimal levels of reading user and sponsored content are then the solutions of the FOC where the marginal utility of reading is equal to the marginal cost:

$$
\begin{aligned}
\alpha^S_1 - \zeta_i - \frac{\alpha^S_2}{K^S_t} r^S_{it} &= \kappa_{1it} + \kappa_{2i} \left(r^S_{it} + r^U_{it} \right) \\
\alpha^U_1 - \zeta_i - \frac{\alpha^U_2}{K^U_t} r^U_{it} &= \kappa_{1it} + \kappa_{2i} \left(r^S_{it} + r^U_{it} \right),
\end{aligned} \tag{A10}
$$

which also implies that the marginal utilities from reading user and site sponsored content are equal. Notice that the marginal utility is the location of an additional post on the line of quality distribution and that a reader will read all the posts whose quality is greater than this post. Therefore, solving for r_{it}^S and r_{it}^U by Equation (A10) is equivalent to selecting the number of highest quality posts from K_t^S and K_t^U combined.

Collecting terms and simplifying, we find that the optimal reading levels for site and user content are given by

$$\begin{bmatrix} r_{it}^{*S} \\ r_{it}^{*U} \end{bmatrix} = \begin{bmatrix} \left(\frac{\alpha_2^S}{K_t^S} + \kappa_{2i}\right) & \kappa_{2i} \\ \kappa_{2i} & \left(\frac{\alpha_2^U}{K_t^U} + \kappa_{2i}\right) \end{bmatrix}^{-1} \begin{bmatrix} \left(\alpha_1^S - \zeta_i - \kappa_{1it}\right) \\ \left(\alpha_1^U - \zeta_i - \kappa_{1it}\right) \end{bmatrix} \tag{A11}$$

This result enables us to conduct a counterfactual of how reading rates and posting rates differ with both the quantity and quality of site sponsored content. Of note, κ_2 captures the substitution effects between the two types of posts. This implies that the increased reading costs induce readers to limit total reading and to choose between the different sources of content.

The ex ante theoretical effect of additional sponsored content on posting is ambiguous. On one hand, there is a competitive effect that lowers the likelihood that users' posts are read, thereby reducing users' incentives to participate in the site. On the other hand, increased content can generate more readership, thereby increasing the utility of posting and the resultant posts.