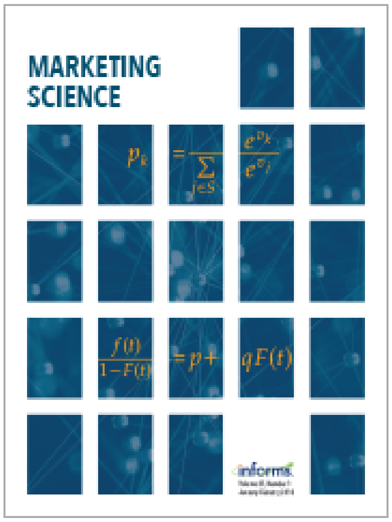


Marketing Science

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

Managing User-Generated Content: A Dynamic Rational Expectations Equilibrium Approach

Dae-Yong Ahn, Jason A. Duan, Carl F. Mela

MARKETING SCIENCE
Vol. 35, No. 2, March–April 2016, pp. 284–303
ISSN 0732-2399 (print) | ISSN 1526-548X (online)

http://dx.doi.org/10.1287/mksc.2015.0937

Managing User-Generated Content: A Dynamic Rational Expectations Equilibrium Approach

Dae-Yong Ahn
College of Business and Economics, Chung-Ang University, Seoul 156-776, Korea, daeyongahn@cau.ac.kr

Jason A. Duan
McCombs School of Business, University of Texas at Austin, Austin, Texas 78712, duanj@mccombs.utexas.edu

Carl F. Mela
Fuqua School of Business, Duke University, Durham, North Carolina 27708, mela@duke.edu

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.

Data, as supplemental material, are available at http://dx.doi.org/10.1287/mksc.2015.0937.

Keywords: user-generated content; marketing strategy; rational expectations; approximate aggregation
History: Received: June 28, 2013; accepted: March 8, 2015; Sanjog Misra served as the guest editor-in-chief and Avi Goldfarb served as associate editor for this article. Published online in *Articles in Advance* November 23, 2015.

1. Introduction

1.1. Overview

By dramatically lowering the cost of content dissemination and consumption, online communication platforms have engendered a rapid proliferation in global user engagement in contexts as diverse as chat rooms, video sharing sites, and Internet forums (e.g., tv.com/forums/ and espn.go.com/nfl/forums). Evidence is afforded by a Google's Ad Planner ranking listing several user sites with substantial user-generated content among the top 20 most trafficked websites (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 (UGC) sites, exceeding $4 billion annually, or more than 10% of firms' online advertising expenditures (eMarketer 2013). Because of their rapid growth, content platforms have seen increasing attention in the literature (Albuquerque et al. 2012, Chevalier and Mayzlin 2006, Dellarocas 2006, Duan et al. 2008, Shriver et al. 2013, Ghose and Han 2011, Huang et al. 2015, Moe and Schweidel 2012, Zhang et al. 2012).

UGC platforms bridge two behaviors, consuming content (e.g., listening or reading) and generating content (e.g., discussing or writing). On the content consumption side of the platform, users generate 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. On the content generation side of the platform, such as posting video game "cheats" and TV show reviews, users obtain 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). As such, users' beliefs about others' current and future participation on the platform are central to the problem of one's own participation, content generation, and consumption. Yet few, if any, papers explicitly consider the potential role of these beliefs on the growth of UGC networks.

Hence, we build on the UGC literature along two key dimensions. First, we allow users' beliefs about others' participation to vary with the state of the network; to this end, we consider a rational expectations equilibrium framework regarding the network's current and future usage. This is material because changes in agents' expectations of the number of users contributing to the site affects whether agents visit the site and consume or create content. Second, we consider dynamic decision making. UGC, like advertising, decays in efficacy over time, analogous to the advertiser problem outlined

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

in Dubé et al. (2005). Because content is somewhat durable (i.e., viewers can find older content to be relevant), there is a trade-off between the marginal cost of investing time in content creation intended for future consumption and creating that content later. Intuitively, one might expect that lower future costs will encourage users to shift content generation to later periods (e.g., when in the office), but this shift must also be weighed against the loss of current period readership. Another source of dynamics arises from the future beliefs about the site's use. If users expect more site participation in the future, they are more likely to participate in the current period. This dynamic regarding future beliefs can accelerate the growth of the network.[2]

These two dimensions lead to a rational expectations dynamic equilibrium wherein each user forms beliefs about many thousands of other users, a task that is cognitively unwieldy for the network's users and computationally infeasible for the researcher. To address these two concerns, we apply and extend the approximate aggregation approach of Lee and Wolpin (2006) and Krusell and Smith (1998).[3] In this approach, users reason that the aggregate evolution of the network should be consistent with the sum of decisions made by all of the members of the network, thereby enabling users to form beliefs about aggregate state transitions in lieu of every other user's actions and state. As a result, the aggregate state transitions across all of the users can vary with changes in the primitives of the system, yielding a structural interpretation of the social engagement problem.

In sum, our contribution is to develop a dynamic, rational expectations equilibrium model of site participation, content generation, and content consumption in the context of a large number of heterogeneous users to shed insight into the management of these networks. We discuss some of these implications next.

1.2. UGC Policy Implications

Enabling users' beliefs to dynamically adapt to changes in market states allows for consideration of policy questions pertaining to the management of UGC platforms. Specifically, we consider how the amount and quality of aggregate content affects users' content generation, content consumption, and likelihood of site participation.

1.2.1. Network Expansion. To increase users' 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 one hand, increased sponsored content attracts more users who can generate more content. In this instance, sponsored content is a strategic complement to user content. On the other hand, sponsored content can dissuade users from generating content, because sponsored and user content are substitutes from the readers' points 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 when the sponsored content crowds out user content.

1.2.2. 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).[4] 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 levels of activity for the hosting platform, often referred to as coordination failure in the economics literature (Liebowitz and Margolis 1994). We consider two means by which to 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 on its most active participants. Marketing strategies such as advertising or incentives and rewards for creating content are not uncommon approaches to attract and retain such users. Second, some network members actively create and consume content, whereas others primarily consume content but rarely, if ever, generate it, a behavior often called "lurking" (Preece et al. 2004). We find that it is not possible to tip a network with just lurkers, although

[2] A more formal characterization of dynamics in our model is presented in §3.7.

[3] Our work builds on the approximate aggregation approach as follows. First, a single unit of supply (e.g., 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 the same 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.

[4] Dubé et al. (2010) and Katz and Shapiro (1994) define tipping as "the degree of market share concentration due to indirect network effects" (Dubé et al. 2010, p. 216). 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.

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 is obtained because lurkers increase overall reading and thus lead to higher content generation utility.

Site-Sponsored Content. An alternative option to tip the network is to substitute SSC for UGC; that is, a firm can sponsor content in an effort to attract content creation and consumption. 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 to do so once the network self-sustains. Such an approach can minimize expenses in the long run, because 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 for tipping to occur.[5] Alternatively, a firm can sponsor content creation at a steady level and thus change users' beliefs about the 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 for tipping to occur. The tipping point is lower than in the first case, because user expectations about increased amount of future content affect current participation decisions. Overall, our results suggest that tipping is feasible with a relatively small amount of high quality sponsored content at the network's inception. This strategy may be quite cost effective, because the expense in creating content manifests only in the early stages of the network. Examples of websites that have pursued this strategy include Soulrider.com (Shriver et al. 2013), a wind-surfing site that jump-started the network by inviting experienced surfers to generate high quality content in its infancy.

1.2.3. Content Quality. In addition to exploring the amount of content, it is also possible to assess the effects of the quality of content users read. We consider two such potential manipulations on the part of the platform: changing the quality of SSC and UGC.

Changing Sponsored Content Quality. For both mature and nascent networks, SSC can have higher average quality than UGC. In our counterfactual analysis, we consider the implications of higher quality content generation strategies by the firm and find that tipping points can be substantially lowered. For the strategy of sponsoring content early on, higher quality can lower the amount of SSC required for tipping from 9.3% to 1%; for the steady SSC strategy, the amount of SSC required for tipping can be lowered from 8% to 1%. Higher quality content can also be especially effective at growing mature networks.

Changing User Content Quality. Sites can also filter user content by eliminating low quality content. 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.

1.2.4. Summary. As noted in §1.1, a key innovation in our model is its equilibrium concept that enables user aggregate beliefs about participation to change dynamically with the states of the system. We find the amount of UGC to tip the network increases from 10.7% to 19.0% of the self-sustaining equilibrium content when updating of beliefs is ignored—an error of nearly 80%. The amount of content needed to tip the network is overestimated in the absence of modeling dynamic expectations because users are not allowed to update their beliefs about potential increases in content in the future. Hence, rational expectations and dynamics play a key role in assessing the effects of policy interventions.

In the balance of this paper, we discuss our data and context and use this information to construct our model. Then we discuss identification and estimation, detail our results, and conduct policy simulations. We conclude with a discussion of key implications and future directions.

2. Data

The data comes from a large Internet site devoted to a common interest, which includes a forum where users can discuss various topics much like fans would discuss a sports team, its players, or various games.[6] However, our model is not restricted to these data; it can be applied more generally to a number of UGC contexts.

2.1. Data Context

Decisions observed in the considered data context include content consumption, content generation, and site participation.

Users *consume content* (such as reading others' posts 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).

Similarly, 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

[5] 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 are diminishing marginal returns to the utility of user content creation.

[6] Because of the context, in the remainder of the text we use *content generation* and *posting* interchangeably, in which case *posting* implies the posting of UGC.

Table 1 Descriptive Statistics

Variable	Mean	Std. dev.	Min.	Max.
Site Participation (n_{it})	0.42	0.49	0	1
Forum Reading (r_{it})	17.97	47.42	0	345
Forum Posting (a_{it})	0.42	1.53	0	19
Individual Post Stock (k_{it})	1.19	2.97	0	19.32

et al. 2004, Nov 2007). There is also a potential negative network externality 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 the actions 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.

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.

2.2. 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.

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 a significant variation in forum reading and posting across individuals as indicated by large standard deviations of these variables. The average individual post stock (defined by Equation (12)) of 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 1(a), 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 1(b) plots the joint distribution (using two-dimensional kernel smoothing) 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 five 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.

2.3. Exploratory Analysis

To explore the potential for indirect network effects and dynamics characterized in §2.1, we conduct a regression analyses using a random data sample of daily activities of 600 users over 61 days.

2.3.1. Individual Reading and Aggregate Post Stock. We regress daily individual reading against aggregate UGC stock (in the unit of 10,000 postings), individual post stock, and an indicator for an end-of-week effect.[7] Table 2 reports that aggregate UGC stock correlates positively with daily individual reading, whereas individual post stock does not correlate with daily reading. The findings suggest the existence of an indirect network effect wherein users tend to read more content when the total amount of content on the site increases.

2.3.2. Individual Posting and Aggregate Reading Rates. Next, we explore the effects of the population's average rate of reading per post on individuals' posting. If users obtain utility from others who read their posts, higher reading rates should increase individuals' utility of posting. Although readers do not directly observe readership counts for every post, we postulate that they can extrapolate their own behaviors to others to develop an expectation of the average reading rate per post. Therefore, we expect individual posting to increase with the average reading rate per post. Moreover, lagged individual posting stock increases the

[7] End of the week is defined as Thursday, Friday, and Saturday, as site usage is lower then. Because site activity is predominantly in the evening (over 50% of reading and posting appears from 4:00 P.M. to 11:59 P.M. Pacific Standard Time) and users tend to be young, we conjecture that the Thursday through Saturday end-of-week drop in site activity is due to the opportunity costs of other types of social activity such as heading out for the evening.

Figure 1 Summary Plots of Reading and Posting

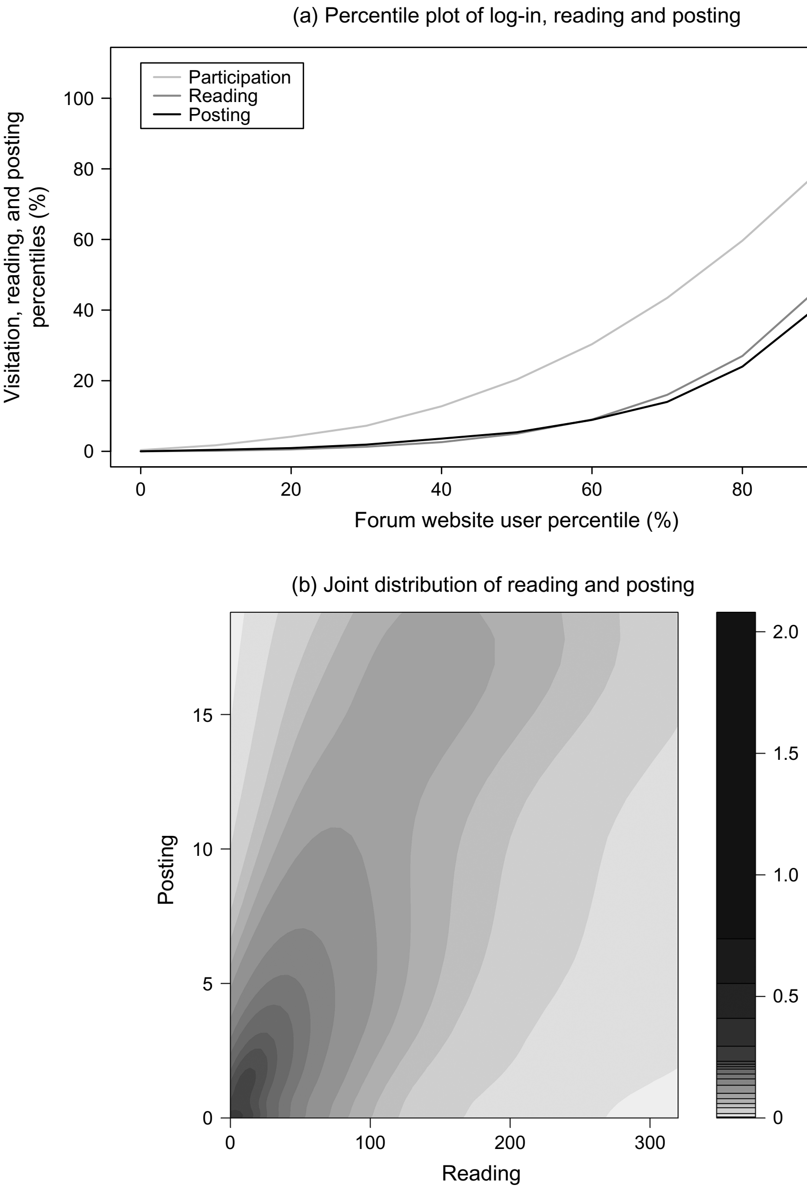


Table 2 The Effect of Aggregate UGC Stock on Individual Reading

Variable	Coefficient (std. error)	*p*-value
Aggregate UGC Stock	0.79 (0.21)	<0.01
Individual Post Stock	0.11 (0.08)	0.19
End-of-Week Effect	−1.03 (0.51)	0.04

current period's entering content, thereby decreasing the marginal utility of content generation. Accordingly, we expect to see a negative correlation between lagged individual post stock and current postings.

As daily individual posts are small integers, we fit a generalized linear model (GLM) using the Poisson family where the daily number of posts is the dependent variable, and the independent variables are listed in Table 3. Table 3 reports the results of this analysis. 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.

Table 3 The Effect of Aggregate Average Reading Rate and Lagged Individual Post Stock on Individual Posting

Variable	Coefficient (std. error)	*p*-value
Average Reading Rate per Post	0.057 (0.026)	0.04
Aggregate UGC Stock	0.12 (0.09)	0.16
Lagged Individual Post Stock	−0.0065 (0.0028)	0.02
End-of-Week Effect	−0.063 (0.002)	0.01

Table 4 **The Effects of Future Average Reading Rate per Post and End of Week on Individual Posting**

Variables	Coefficient (std. error)	p-value
Average Reading Rate per Post (current period)	0.14 (0.035)	<0.01
Average Reading Rate per Post (next period)	−0.053 (0.024)	0.03
Aggregate UGC Stock	0.0048 (0.0071)	0.50
Lagged Individual Post Stock	−0.013 (0.0031)	<0.01
End-of-Week	−0.051 (0.0024)	0.03
Day Before End-of-Week	0.11 (0.026)	<0.01

2.3.3. Exploratory Evidence of Dynamic Behavior. Following the approach outlined in Chintagunta et al. (2012), we conduct exploratory analyses of dynamic behavior. Chintagunta et al.'s (2012) approach 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 at the end of the week, as evidenced by the large decrease in posts observed then (perhaps because of the increased opportunity costs of leisure time such as evenings out on Thursday, Friday, and Saturday). Hence, increased content generation prior to the end of the week suggests users strategically manage posts to avoid higher future posting costs. Likewise, because users can create costly content today or tomorrow for consumption tomorrow, there may be an incentive to delay content creation to when its consumption will be higher.

We analyze these timing decisions by fitting the GLM with the Poisson family for the daily number of postings.[8] The results in Table 4 indicate a negative effect (p-value < 0.05) of the reading rate per post of the next period on current period postings, suggesting that users delay content generation until there is an audience available to read it. A similar effect is evidenced for the impending end of the week; the indicator of the day before the end of the week is positive (p-value < 0.01), implying that users move content generation before the end of the week when posting costs are higher.[9]

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 §3, discussed next.

3. Model

This section formalizes the model for the decisions outlined in §2.1. First, we consider content consumption in the face of heterogeneous quality of UGC in §3.1. Second, we outline the role of consumption on content generation in §3.2. Third, we discuss site participation conditional on these two decisions in §3.4. Finally, we discuss users' strategic behaviors in content generation implied by the dynamic model in §3.7.

In sum, we consider (i) a population of M users' decisions (indexed by $i = 1, \ldots, M$) to participate on a content sharing website, $n_{it} = \{0, 1\}$, 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, first solving for the content consumption and generation decisions, and then solving for the participation decision.

3.1. Content Consumption

3.1.1. Reading Utility. Readers consume content when the benefit exceeds costs. The utility of reading is dependent on 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 $Q_1, \ldots, Q_{K_t}$, are independent and identically distributed (i.i.d.) Uniform$[L, U]$.[10]

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

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

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 can be expressed as

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

[8] We note the following analyses are exploratory because there are no formal statistical tests of dynamics (Manski 1993, Magnac and Thesmar 2002). The relationship to the dynamics in our model is motivated in Web Appendix D.3 (available as supplemental material at http://dx.doi.org/10.1287/mksc.2015.0937).

[9] For a robustness check, we also test three additional models by eliminating the effects of day before the end of the week (Model 2) and the end of week (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.

[10] In §6 we generalize our analysis to allow for SSC to differ in its quality from UGC.

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

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

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

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

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 $[K_t + (\alpha_1-\alpha_2)/\alpha_1 + \alpha_2/(2\alpha_1)]/(K_t+1) \approx 1$. Thus, reader i's utility of reading becomes

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

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). Note that 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 of 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

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

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

3.1.2. 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

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

where the cost function is convex when κ_{1it} and κ_{2i} are both positive. Cyclicality, such as the effect of the end of the week, 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 an indicator for the end of the week. 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

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

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

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

Because of heterogeneity in reading costs across segments, r_{it}^* differs across segments. After user i decides to visit the website, she realizes 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 $r_{it} = r_{it}^* \nu_{it}$.[11] Therefore, ex post amount of reading will also differ across users in the same segment. Because 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 the posts.

3.2. Content Generation

3.2.1. 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).[12] This expected rate of reading per post, y_t, is defined by

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

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

[11] Because ν_{it} is realized after a user visits the site, it is independent of the 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 Web Appendix C.1 to facilitate the maximum likelihood estimation.

[12] In Web 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 estimation. However, it is necessary to recompute this rational expectation equilibrium in our counterfactual analyses. (See §3.6 for more details.)

content. This constitutes an indirect network effect on posting from reading. Second, there is a competitive effect of K_t in the denominator because 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.

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}, respectively, 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 utility function with diminishing marginal return is

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

3.2.2. 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 four 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 that these costs vary exogenously with the day of the week.

The cost of posting is specified as

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

where the random error in the cost function, $\varepsilon_{it}(a_{it})$, has a generalized extreme value distribution. The time-invariant component of the linear marginal cost ξ_i, which is heterogeneous across users, is assumed to follow a latent segment structure. In addition, τ_{it} models a cyclical effect such as letting $\tau_{it} = \tau_i w_t$, where w_t is an end-of-week 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.

3.2.3. Optimal UGC Generation. We presume a user chooses the optimal number of postings a_{it}^* (amount of content to generate) to maximize the expected discounted sum of per-period utilities minus per-period costs

$$u^P(a_{it}) - c_{it}^P(a_{it}) + E\left\{\sum_{h=t+1}^{\infty} \beta^{h-t}[u^P(a_{ih}) - c_{ih}^P(a_{ih})]\right\}. \tag{15}$$

This summed utility can be formulated as a dynamic programming problem with the Bellman equation

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

where the $\bar{c}_{it}^P(a_{it}) = (\tau_{it} + \xi_i)a_{it}$ represents the nonrandom component of the posting costs. The value function $V_i(s_{it}, \varepsilon_{it})$ depends on $s_{it} = \{k_{i,\,t-1}, K_t, w_t\}$ and ε_{it}, which are the state variables. The number of per-period postings a_{it} is the control variable, which is discretized because over 99% of users create fewer than 10 posts a day. Hence, we let $a_{it} \in A = \{0, 1, 2, \ldots, \bar{a}\}$, where $\bar{a}$ represents the maximum number of daily postings. Define the integrated value function $\widetilde{EV}_i(s_{it}, a_{it})$ as

$$\begin{aligned} \widetilde{EV}_i(s_{it}, a_{it}) = \int_{s_{i,\,t+1}} \int_{\varepsilon_{i,\,t+1}} &V_i(s_{i,\,t+1}, \varepsilon_{i,\,t+1}) \\ &\cdot p(s_{i,\,t+1}, \varepsilon_{i,\,t+1} \mid s_{it}, \varepsilon_{it}, a_{it})\, ds_{i,\,t+1}\, d\varepsilon_{i,\,t+1}. \end{aligned}$$

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

$$\begin{aligned} &P(a_{it} \mid s_{it}, n_{it} = 1) \\ &\quad = \frac{\exp(u^P(a_{it}) - \bar{c}_{it}^P(a_{it}) + \beta\widetilde{EV}_i(s_{it}, a_{it}))}{\sum_{a'_{it}\in A}\exp(u^P(a'_{it}) - \bar{c}_{it}^P(a'_{it}) + \beta\widetilde{EV}_i(s_{it}, a'_{it}))}. \end{aligned} \tag{17}$$

3.3. Modeling Heterogeneity in UGC Consumption and Generation

We jointly model the heterogeneity parameters in the reading and posting models with J latent segments

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

In Equation (18), $\bar{\zeta}_j$, $\bar{\kappa}_{1j}$, $\bar{\kappa}_{2j}$, $\bar{\xi}_j$, and $\bar{\tau}_j$ are the segment-specific values for the parameters in the content consumption and generation models.

Because reading and posting costs follow a distribution that is jointly estimated, our model captures a wide array of interdependent behaviors for 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 positively correlated across users.

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

3.4. Site Participation

Prior to posting and reading, a user must decide whether to visit the UGC website, and this decision is predicated on 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

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

where μ_1 and η are scale parameters that rescale the utility of reading and the contextual shock relative to the utility of posting, respectively. 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 their previous posting stock, given by $u^P(\rho k_{i,t-1} y_t) + \beta E\tilde{V}_i(s_{it}, a_{it}=0)$. 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

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

A user chooses to visit the website 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 i.i.d. 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

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

Then the choice probability of visiting the site can be written as

$$\begin{aligned} &P(n_{it}=1 \mid s_{it}) \\ &= \left(\exp\left\{\mu_1 E \max_{r_{it}} [u^R(r_{it}) - c_{it}^R(r_{it})] + \eta IV_{it}\right\}\right) \\ &\quad \cdot \left(\exp\{\mu_{0j} + \eta[u^P(k_{i,t-1} y_t) + \beta E\tilde{V}_i(s_{it}, a_{it}=0)]\}\right. \\ &\quad \left. + \exp\left\{\mu_1 E \max_{r_{it}} [u^R(r_{it}) - c_{it}^R(r_{it})] + \eta IV_{it}\right\}\right)^{-1}, \end{aligned} \tag{22}$$

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

3.5. 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, evolve deterministically over time. Last, 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}; $i = 1, \dots, M$. However, we assume that from the perspective of any individual user, K_t evolves stochastically given K_{t-1}, but independent of the individual's 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 the other users, nor believe her own action a_{it} has any influence on the aggregate UGC stock, K_t. If we impose the rational expectations constraint, then user i's belief about the state transition for K_t must coincide with the actual behaviors of all users on the site, as discussed next.

3.6. Rational Expectations and Approximate Aggregation

Rational expectations require that users' beliefs about the evolution of K_t be consistent with the actual transition, which is the sum of all individuals' posting decisions. This observation becomes germane 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. Accordingly, users' beliefs can change in response to a change in the strategy of the site.

3.6.1. 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 end-of-week indicator, and ω_2^K represents the 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.[13] From an individual's perspective, there is a degree of uncertainty 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 as follows:

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

where ω_2^y is again for the effect for the end of the week. 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 approximate the expected one.[14] See Web Appendix A for details. The parameters ω_0^y, ω_1^y, and ω_2^y are also determined by the rational expectations equilibrium.

3.6.2. Consistency of Approximate Aggregation.

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

$$K_t = \rho K_{t-1} + \sum_{i=1}^{M} a_{it}^*. \tag{25}$$

Using Equation (25) directly to calculate users' rational expectations requires us to assume that every user knows how each and every other user chooses their optimal action a_{it}^* given their respective policy functions and states. Complete knowledge of the optimal behavior of thousands of other users is an unrealistic assumption that imposes a large informational burden on individual users. 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 optimal behaviors n_{it}^*, a_{it}^*, and r_{it}^*. Aggregating across users, we recompute K_t and y_t and recompute individual behaviors, iterating back and forth between the individual-level and aggregate models until convergence. Web Appendix B details the algorithm used to simulate a rational expectations equilibrium. The parameters ω_0^K, ω_1^K, ω_2^K and ω_0^y, ω_1^y, ω_2^y are reestimated in every step of the iterations to find the fixed point of the rational expectations equilibrium.[15] 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 rational expectations equilibrium with approximate aggregation by simulation in Web Appendix D.

3.7. 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, the cost function as linear, and abstract away from the error (we suppress individual index i for clarity), we obtain

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

The Euler equation from this simplified problem can be derived 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 trade-off 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

[13] Note that the order of the state transition equations cannot be higher than the order of the individual-level model, or else the individual level model would fail to account for a users' beliefs about these higher order states. Here we assume that 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. Because 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 §2.3, we test and find the AR(1) model is the best model for our data.

[14] We also approximate 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 that the estimation results remain unchanged, so the approximation by Equation (24) is robust.

[15] In estimation, the aggregate states are observed (reflecting the current equilibrium), so no iteration to reestimate ω_0^K, ω_1^K, ω_2^K and ω_0^y, ω_1^y, ω_2^y is necessary. In the counterfactual analyses, states need to be computed.

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 the 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 involve shifting content creation to align with periods with higher reading rate y_t and lower posting cost τ_t. Finally, we note that, all else equal, an increase in entering posting stock, k_{t-1}, implies an attendant reduction in a_t to maintain the equality, k_t. The rationale behind this relationship is that there is an optimal level of current and past content that can be obtained either by having a larger stock of old content or an increase in current content.

Next, we consider the role of expected network effects in dynamics as discussed in §3.6. 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 availability 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.

4. Estimation and Identification

4.1. Estimation

Within each iteration of the likelihood optimization algorithm, an efficient estimation approach using maximum likelihood (MLE) 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 ω_0^K, ω_1^K, ω_2^K and ω_0^y, ω_1^y, ω_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 reestimate ω_0^K, ω_1^K, ω_2^K and ω_0^y, ω_1^y, ω_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 by MLE, using the joint likelihood function of reading and posting for each individual user i

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

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

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

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. Su and Judd (2010) show that the two-step pseudo-maximum likelihood estimator discussed above is consistent. We bootstrap to compute the standard errors. See Web Appendix C.2 for details.

4.2. Identification

4.2.1. Post Stock Decay Parameter ρ. As indicated in §3.2, 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, the aggregate UGC stock, and the average participation rate are stationary over 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.[16]

4.2.2. 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 r^*_{it} for person i at time t is equal to $(\alpha_1 - \bar{\zeta}_j - \bar{\kappa}_{1j} w_t)/(\alpha_2/K_t + \bar{\kappa}_{2j})$ if person i belongs to 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 posting, $\bar{\kappa}_{1j}$ and $\bar{\tau}_j$, we use an indicator variable that is set to 1 for the end of the week and 0 otherwise. Hence, these parameters are identified by differences in the mean amount of reading and number of postings between the end of the week and the other days in the week. 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 (e.g., Manski 1993, Magnac and Thesmar 2002). Hence we set $\beta = 0.99$.[17] Finally, there are scale parameters 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.

5. Results

5.1. Decay Parameter and Initialization of Post Stock

As indicated in §3.2, the decay rate ρ quantifies the rate of diminution in posting stock over time. This decay rate is estimated using an auxiliary data set 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 data set records the daily number of times that these posts are read over 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 in the t-th period after it is posted. We estimate ρ using the model $z_{kt} \sim \text{Poisson}(\rho^{t-1}\lambda_k)$, where λ_k captures the heterogeneity in the amount of reading among posts. The MLE using generalized linear models yields $\widehat{\log \rho} = -0.30$ (0.0036), so the decay parameter estimate is 0.74.

Because there is no history of posts prior to the initial week, individuals' initial post stocks 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 until the individual post stock reaches a steady state. The individual's steady state is then reused 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 reached the steady state at the starting period of our data. The aggregate stock, K_t, is computed by aggregating individual stocks.

5.2. Aggregate Variables Under Rational Expectations

Section 3.6 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 autocorrelation ($\omega_1^K = 0.93$) for the aggregate UGC stock. The average reading rate per post is an increasing function of aggregate UGC stock ($\omega_1^y = 0.054$ for UGC in unit of 10,000 postings), which implies a

[16] 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 §5.1), 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.

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

Table 5 Estimation Results for the Aggregate UGC Stock Transition Equation and Average Reading Rate per Post Equation

Model parameter	AR(1) for UGC stock	Reading rate per post
Intercept ω_0^K or ω_0^y	2.89 (1.35)	6.02 (0.94)
Lagged aggregate UGC stock ω_1^K	0.93 (0.33)	—
Current aggregate UGC stock ω_1^y	—	0.054 (0.0052)
End-of-week effect ω_2^K or ω_2^y	−0.69 (0.094)	−0.55 (0.035)
Residual R^2	0.89	0.51

Notes. Standard errors are in parentheses. All parameters are significant at p-value < 0.05.

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

Parameters	First segment; frequent users	Second segment; light users
Content generation model		
Utility coefficient γ	0.79 (0.015)	
Cost coefficient $\bar{\xi}_j$	1.29 (0.087)	6.83 (0.33)
End-of-week effect $\bar{\tau}_j$	8.99 (3.93)	9.60 (4.31)
Content consumption model		
Utility coefficient α_2	0.60 (0.092)	
Utility heterogeneity coefficient $\bar{\zeta}_j$	−0.70 (0.082)	0.70 (0.082)
End-of-week cost effect $\bar{\kappa}_{1j}$	−0.049 (0.040)	0.0052 (0.014)
Quadratic cost coefficient $\bar{\kappa}_{2j}$	0.026 (0.0023)	0.011 (0.0021)
Site participation model		
Intercept, μ_{0j}	27.12 (13.23)	4.46 (2.11)
Reading scale parameter μ_1	0.51 (0.23)	
Gumbel scale parameter η	0.97 (0.014)	
Heterogeneity		
Segment size	43% (3.1%)	57% (3.1%)

Note. Bootstrapped standard errors are in parentheses.

positive network effect of posting. The effect of the end of a week in model (24) is significantly negative, which means lower reading activity at the end of the week.

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 $= 0.73$). The 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 equal to 0.63, not rejecting the null hypothesis that the autocorrelation of the residuals is 0. Therefore, our data support the assumption that users can use approximate aggregation and AR(1) to predict K_t on the basis of rational expectations.

5.3. 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 1 to November 30, 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.[18]

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, which reflects the normalization needed for identification, as we discuss in §3.3.

Comparing the two groups, the second segment is larger in size and evidences higher reading and posting costs; hence, this group of users reads less often and rarely posts content. Thus, we denote them "light users." Also of note, the effect $\bar{\tau}_j$ in the posting cost function is positive, so the users tend to post less at the end of the week.

Next we consider the parameter estimates in the site participation model. As indicated in Table 6, the first 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 significantly between segments in the site participation model. This suggests that heterogeneity in site participation decisions are largely predicated on reading and posting utilities.

To quantify model fit, we simulate user reading, posting, and visiting data given the estimated parameter values in Table 6. We then calculate the mean absolute percentage error (MAPE) between the simulated aggregate UGC, the average reading rate per post, and the number of visitors across 61 observed days and the corresponding observed values. These values are, respectively, 7.59%, 4.47%, and 7.04%, which shows that the model fits the data well.[19]

[18] We also test three segments of users. However, the Bayesian information criterion for the three-segment model is higher than that for the two-segment model.

[19] 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 intertemporal ability to choose when to post. We find this static model fits the data much more poorly (log-likelihood $= -1.50774 \times 10^5$) than our dynamic model (log-likelihood $= -9.9972 \times 10^{-4}$). In the second model, we aggregate data to the weekly level to filter out daily variation and estimate a static model. Because 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.

6. 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 SSC. We consider SSC that looks identical to UGC from the perspective of the reader except in their quality (e.g., the sponsorship is not revealed in essence disguising SSC as UGC). We focus on disguised 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 for user and sponsored content, r_{it}^S and r_{it}^U, are the solutions to the optimization problem of the total reading payoff. In Web Appendix E, we derive the optimal total amount of reading and how it is allocated between UGC and SSC given by

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

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.

6.1. 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. Note 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 are plotted in Figure 2.

Figure 2 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 a sufficiently large quantity of much higher quality SSC ($\alpha_2^S/\alpha_2^U = 10\%$), the site can increase the participation rate by as much as 25%. Hence, the sponsored content arc elasticity is about 0.5 and 2.5 for these respective cases.

The increase in UGC is much less pronounced than the increase in participation owing to the competition between UGC and SSC. For a 10% increase in high quality SSC, there is a 6% increase in aggregate UGC (i.e., an arc elasticity equal to 0.6). For a 10% increase in equal quality SSC, there is a 1% increase in UGC (i.e., an arc elasticity equal to 0.1). An increase in high quality SSC from 7% to 10% actually dampens UGC, presumably because user posts are affected by the shift of readers to higher quality site-sponsored posts.

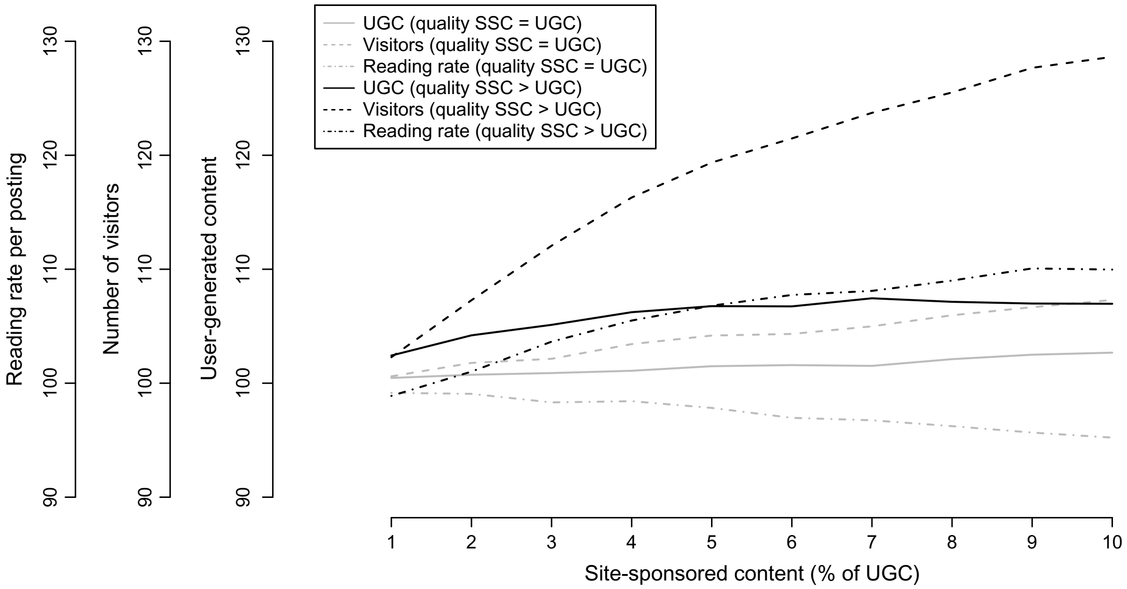
Similarly, the reading rate per post decreases to about 96% in the case when UGC and SSC have equal quality because of the competition effect. In the high quality SSC case, the increase in reading rate per post is moderate at about 9%.

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

6.2. 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 declined, and the domain name has been sold to TV.com.

Figure 2 **Effect of SSC Strategies on Increasing Site Traffic**



Notes. 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.

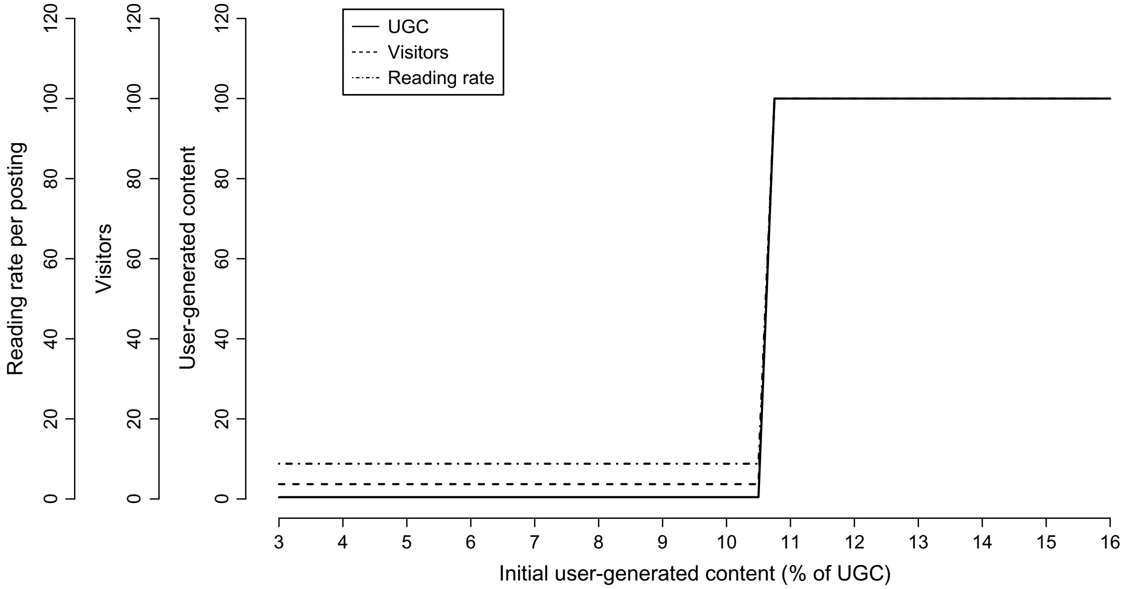
To prevent network contraction, one way to tip the network to a self-sustaining high-activity equilibrium is to attract *user* content to the site (e.g., via marketing or advertising). A second approach, such as the strategy employed by Soulrider.com, is to sponsor a sufficiently high level of content 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. Although a definitive answer to these three strategies (incentivizing initial UGC, creating initial sponsored content only or regular sponsored content) depends on 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.

6.2.1. UGC on Tipping. We first consider the role of UGC stock on network tipping. Figure 3 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%. By 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. Stated differently, the dynamics in the system, coupled with rational expectations, play a major role in the tipping point. When beliefs about the participation of others is not allowed to evolve with changes in the system, we find that 19% of the current observed levels of UGC is needed to tip the network. The rationale behind the 78% increase in the tipping point value when rational expectations are ignored is provided in §3.7; the updated beliefs about higher future reading rates and content increases the incentives to participate in it in the current period, thereby lowering the tipping point.

Next we consider the role of user heterogeneity in tipping. 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 alone cannot easily tip the network. However, they can have a considerable indirect effect on tipping, because 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

Figure 3 The Critical Point of Initial UGC



Notes. 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.

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 to increase the marginal impact of heavy users on tipping.

Because heavy users generate more content, one might surmise that targeting this group would tip the network the fastest. Indeed, for the frequent user segment, an initial endowment of the user post stock at 11% of the high-activity equilibrium level can tip the network even if we set the initial stock of light users to zero. By 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 frequent users to zero. Hence, the frequent user segment's role in network tipping is far more substantial, but a large population of lurkers can also tip a network.

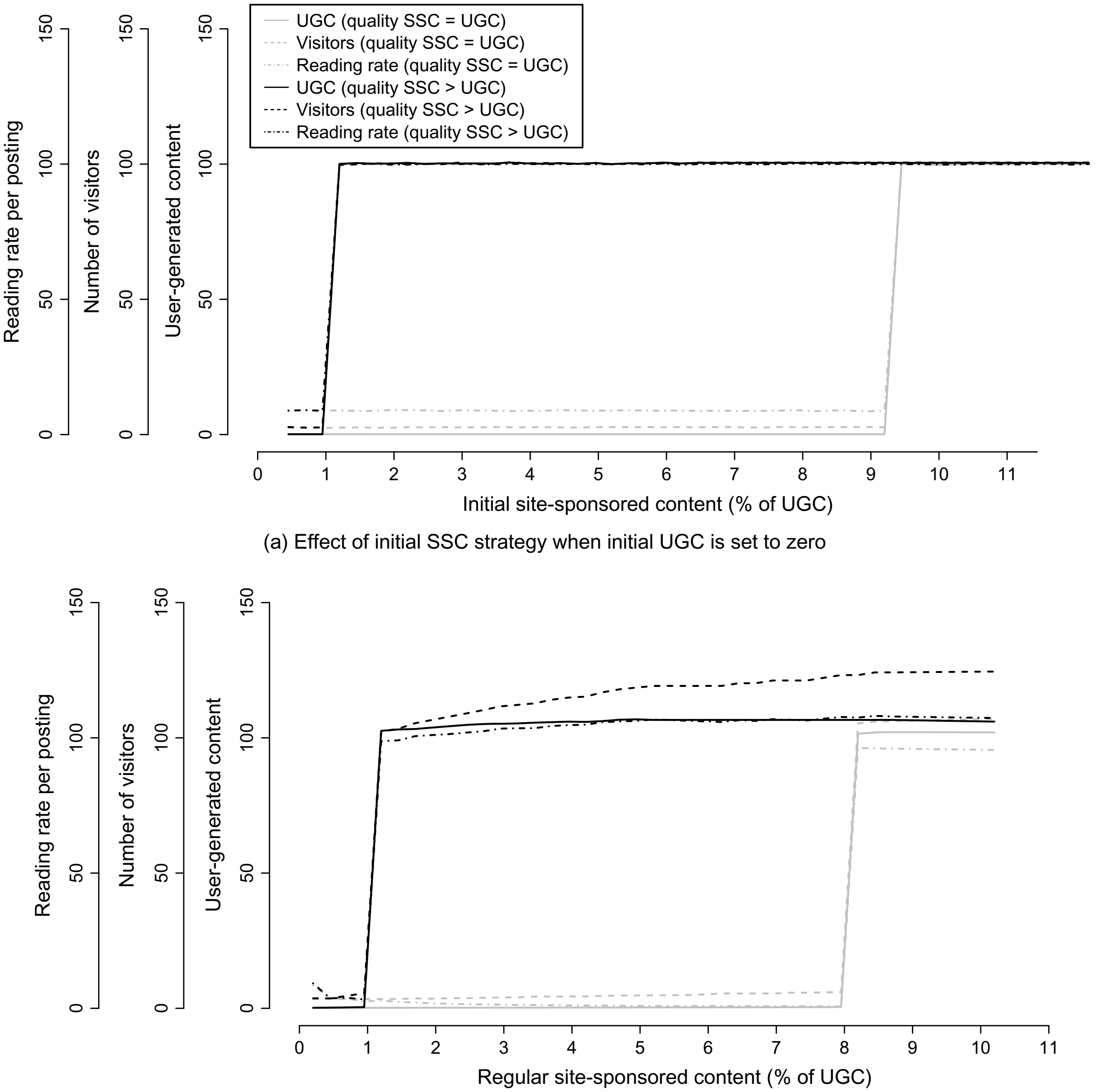
6.2.2. SSC on Tipping. Next we consider the role of SSC on tipping; that is, the site may invite sponsored content to "jump-start" user activity. We consider two approaches to jump-start the network, In the first, the site increases the initial SSC only. This strategy is analogous to sponsoring posts 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 these analyses are reported in Figure 4.

Results shown in Figure 4(a) suggest that the initial SSC of 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 4(b), the effect of sponsoring content on a 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 arises because users perceive a regular stream of high quality content availability. Hence, we find it possible for the site to tip the network with a relatively small number of high quality posts, so long as the site is committed to sponsoring content for an indefinite period of time.

In summary, we contrast three strategies: (i) enlisting 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

Figure 4 **Effect of Sponsored Content Strategies**



(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

Notes. The horizontal axes represent initial and/or regular SSC as a percentage of the UGC in the high-activity equilibrium. The vertical axis depicts the steady-state site activity measures.

decreases. If the site sponsors posts instead, the value of SSC remains high enough to attract readers, but the individual user stock is sufficiently low, enhancing their incentive to post. Because of this, the site tips more quickly.

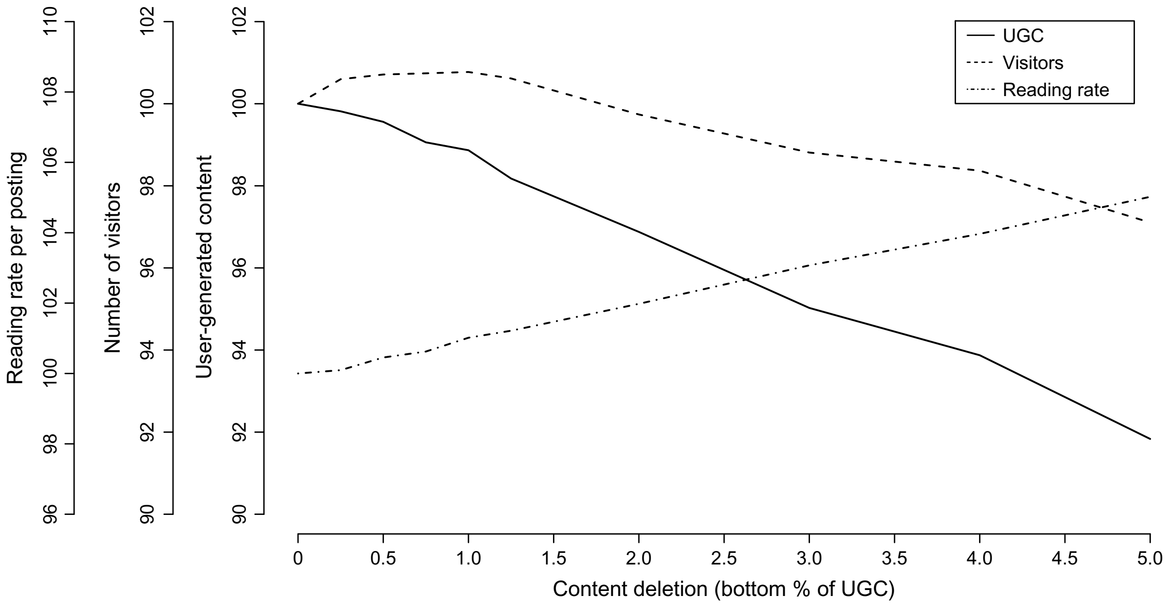
The preferred strategy would be incumbent on the relative cost of each. It is more effective for the site to jump-start with just initial posts (Figure 4(a)) than with a constant stream of posts (Figure 4(b)). However, the cost of the latter strategy is borne each period rather than once. Were the site 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 a profound effect on networks taking off, but a rather limited effect on the current steady-state level of user engagement.

6.3. Managing UGC Quality

Finally, we consider the possibility of filtering user content by removing low quality posts such as profanity or trolling. The results of this counterfactual are presented in Figure 5.

Predictably, reading rates increase as average quality increases. However, knowing that content is subject to removal, users tend to generate less content on

Figure 5 Managing UGC Quality



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 by the site. 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%.

With a static utility model, the improvement in site use is overestimated by one percentage point when 1% of content is filtered, and the decrease in site use is underestimated by six percentage points when 5% of content is filtered. In other words, visitation is consistently overestimated. We conjecture this bias arises because the posting utility loss from not having future content not read is ignored; hence, site participation utility increases.

7. Conclusions

Recent advances in technology and media have enabled UGC sites to become an increasingly prevalent source of information for consumers as well as a more relevant channel for advertisers to reach users of these sites. Hence, the factors driving the use of these networks is of 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 on 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 others' reading the posts. Underpinning these two behaviors are users' beliefs regarding how the aggregate amounts 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 on whether initial usage can cross a sufficient threshold to attract participation. Another theoretical insight is that UGC and SSC 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. Finally, we note that dynamics play a role in determining the tipping. Ignoring future value of site participation leads to a lower utility of engagement and, therefore, higher tipping thresholds.

In 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 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. When sponsored post quality is sufficiently high, it only takes about 1% of the observed posting levels in the mature network for tipping to occur. Our study offers evidence of the efficacy of this strategy. Finally, we consider the impact of filtering low quality user content. 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.

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. Our analysis focuses on the user side and abstracts from modeling the payoffs of UGC websites. This might be reasonable in many UGC contexts because studying dynamics and network effects on one site is important, but the competition between platforms is an area of interest. Second, it would be useful to extend our model to capture heterogeneity in content information to explore what information is most relevant in increasing site engagement. Relevantly, 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 the requirement for numerical computation to become feasible. Third, thanks to the institutional details characterizing our problem, it is not amenable to a closed form solution. An analytic approach might yield more generalized insights. 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 UGC and the application of the rational expectations equilibrium with approximate aggregation in marketing.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/mksc.2015.0937.

Acknowledgments

The authors would like to thank Joel Huber, Wagner Kamakura, Vineet Kumar, Oded Netzer, and seminar participants at Carnegie Mellon University, Erasmus University, Indiana University, King Carlos III University, Tilburg University, University of Chicago, University of Texas, and the 2011 Quantitative Marketing and Economics Conference for their comments. The authors also thank the Networks, Electronic Commerce and Telecommunications (NET) Institute (www.netinst.org) for partial financial support of this project.

References

Albuquerque P, Pavlidis P, Chatow U, Chen KY, Jamal Z (2012) Evaluating promotional activities in an online two-sided market of user-generated content. *Marketing Sci.* 31(3):406–432.

Bughin JR (2007) How companies can make the most of usergenerated content. *McKinsey Quart.* 1–4.

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

Chintagunta PK, Goettler RL, Kim M (2012) New drug diffusion when forward-looking physicians learn from patient feedback and detailing. *J. Marketing Res.* 49(6):807–821.

Clarke DG (1976) Econometric measurement of the duration of advertising effect on sales. *J. Marketing Res.* 13(4):345–357.

Dellarocas C (2006) Strategic manipulation of Internet opinion forums: Implications for consumers and firms. *Management Sci.* 52(10):1577–1593.

Duan W, Gu B, Whinston AB (2008) Do online reviews matter?—An empirical investigation of panel data. *Decision Support Systems* 45(4):1007–1016.

Dubé JP, Fox JT, Su CL (2012) Improving the numerical performance of static and dynamic aggregate discrete choice random coefficients demand estimation. *Econometrica* 80(5):2231–2267.

Dubé JP, Hitsch GJ, Chintagunta PK (2010) Tipping and concentration in markets with indirect network effects. *Marketing Sci.* 29(2): 216–249.

Dubé JP, Hitsch GJ, Manchanda P (2005) An empirical model of advertising dynamics. *Quant. Marketing Econom.* 3(2):107–144.

eMarketer (2013) eMarketer in review–Key 2013 trends, coverage areas and platform growth. http://www.emarketer.com/newsroom/index.php/emarketer-review-key-2013-trends-coverage-areas-platform-growth/.

Ghose A, Han SP (2011) A dynamic structural model of user learning on the mobile Internet. Working paper, New York University, New York.

Ghosh A, McAfee P (2011) Incentivizing high-quality user-generated content. *Proc. 20th Internat. Conf. World Wide Web* (ACM, New York), 137–146.

Hennig-Thurau T, Gwinner KP, Walsh G, Gremler DD (2004) Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet? *J. Interactive Marketing* 18(1):38–52.

Huang Y, Singh PV, Ghose A (2015) A structural model of employee behavioral dynamics in enterprise social media. *Management Sci.* 61(12):2825–2844.

Karat CM, Halverson C, Horn D, Karat J (1999) Patterns of entry and correction in large vocabulary continuous speech recognition systems. *Proc. SIGCHI Conf. Human Factors Comput. Systems* (ACM, New York), 568–575.

Katz ML, Shapiro C (1994) Systems competition and network effects. *J. Econom. Perspectives* 8(2):93–115.

Katz ML, Shapiro C (1998) Antitrust in software markets. Eisenbach JA, Lenard TM, eds. *Competition, Innovation and the Microsoft Monopoly*: *Antitrust in the Digital Marketplace* (Kluwer Academic Publishers, Boston), 29–81.

Krusell P, Smith AA (1998) Income and wealth heterogeneity in the macroeconomy. *J. Political Econom.* 106(5):867–896.

Lee D, Wolpin KI (2006) Intersectoral labor mobility and the growth of the service sector. *Econometrica* 74(1):1–46.

Liebowitz SJ, Margolis SE (1994) Network externality: An uncommon tragedy. *J. Econom. Perspectives* 8(2):133–150.

Magnac T, Thesmar D (2002) Identifying dynamic discrete decision processes. *Econometrica* 70(2):801–816.

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

Mela CF, Gupta S, Lehmann DR (1997) The long-term impact of promotion and advertising on consumer brand choice. *J. Marketing Res.* 34(2):248–261.

Moe WW, Schweidel DA (2012) Online product opinions: Incidence, evaluation, and evolution. *Marketing Sci.* 31(3):372–386.

Nardi BA, Schiano DJ, Gumbrecht M, Swartz L (2004) Why we blog. *Comm. ACM* 47(12):41–46.

Nov O (2007) What motivates Wikipedians? *Comm. ACM* 50(11):60–64.

Preece J, Nonnecke B, Andrews D (2004) The top five reasons for lurking: Improving community experiences for everyone. *Comput. Human Behav.* 20(2):201–223.

Ransbotham S, Kane GC, Lurie NH (2012) Network characteristics and the value of collaborative user-generated content. *Marketing Sci.* 31(3):387–405.

Rust J (1994) Structural estimation of Markov decision processes. Engle RF, McFadden DL, eds. *Handbook of Econometrics*, Vol. 4 (North-Holland, Amsterdam), 3081–3143.

Ryan SP, Tucker C (2012) Heterogeneity and the dynamics of technology adoption. *Quant. Marketing Econom.* 10(1):63–109.

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

Stigler GJ (1961) The economics of information. *J. Political Econ.* 69(3):213–225.

Su CL, Judd KL (2010) Structural estimation of discrete-choice games of incomplete information with multiple equilibria. *Proc. Behavioral Quant. Game Theory: Conf. Future Directions, BQGT '10* (ACM, New York), Article 39.

Zhang K, Evgeniou T, Padmanabhan V, Richard E (2012) Content contributor management and network effects in a UGC environment. *Marketing Sci.* 31(3):433–447.