Managing User Generated Content

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

June 19, 2013

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

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

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

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

Abstract

This paper considers the creation and consumption of content on user generated content (UGC) platforms, e.g., reviews, articles, chat, videos, etc. On these platforms, users' expectations regarding the participation of others on the site become 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 site sponsored content (SSC) strategies. We find sponsoring content can be effective at creating a self-sustaining network, but once the network tips, sponsored content does little to increase usage.

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

1 Introduction

By dramatically lowering the cost of content dissemination and consumption, online communication platforms have engendered a rapid proliferation in global user engagement. Evidence is afforded by a recent ranking done by Google's Ad Planner, listing several user sites with substantial user generated content among the top 20 most trafficked web sites (`YouTube.com`, `Wikipedia.com`, `Mozilla.com`, `Wordpress.com`, `Ask.com`, `Amazon.com` and `Taobao.com`).[1] Coincident with this increase, advertisers are spending more of their budget on social media and user generated content sites (UGC), exceeding $2BB annually, or more than 8% of firms online advertising expenditures (eMarketer 2010).

UGC platforms rely upon two behaviors, consuming content (e.g., listening or reading) and generating content (e.g., discussing or writing). Content consumption generates utility via the pleasure of reading or the utility of information. Content generation, like posting video game "cheats" and TV show reviews, yields utility from the reputation effect of being influential, knowledgeable or popular, suggesting utility increases as more content is consumed (Bughin 2007; Hennig-Thurau et al. 2004; Moe and Schweidel 2012; Nardi et al. 2004; Nov 2007).[2] Accordingly, the content generation decision is predicated on beliefs about the number of other people consuming and generating content (Shriver et al. 2013). As such, users' beliefs about others' participation on the platform are central to the problem of content generation and consumption. In spite of this few, if any papers, explicitly consider the role of these beliefs on the growth of UGC networks.

We address this gap by capturing the evolution in beliefs about future consumption and generation of content; that is, we allow these beliefs about the site participation of others to be endogenous, leading to a dynamic rational expectations equilibrium model of user generated content and consumption in the context of heterogeneous users. This rational expectations equilibrium forms the basis of a joint dynamic model of content consumption and generation. Owing to its structural orientation, this approach enables us to address a number of policy questions of interest to UGC platforms.

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

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

- Content generation in a mature network. To increase consumers' utility of consumption, platforms can provide more site sponsored content (SSC); for example, an online forum site could invite experts to create additional content to supplement that of users. These expert posts are site sponsored, but can be made indistinguishable from UGC from the reader's perspective except in their quality. The problem of managing this type of SSC is challenging. On the one hand, increased sponsored content attracts more users who will likely to post more content, because of the increased availability of information. In this instance, sponsored content is a strategic complement to user content. On the other hand, sponsored content can dissuade users from posting content because sponsored and user content are substitutes from the reader's point of view. The optimal amount of sponsored content, therefore, becomes a question of the relative magnitude of these various effects. In our context, sponsored and user content are strategic complements at low levels of sponsored content, but become substitutes as the sponsored content crowds the user content.

- Content generation in a new network. 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). Owing to the dynamics in beliefs, early content can have a profound effect upon whether the network grows or implodes. Without sufficient reading mass, content generators might believe there is little value in creating content, thereby causing the network to become ensnared in an undesirable equilibrium with very low level of activity for the hosting platform; often referred to as network failure in the economics literature(Liebowitz and Margolis 1994). Related, the concept of self-fulfilling prophecies are germane in the context of social engagement, because the beliefs that others will enter the site can induce a herding behavior towards using the site. Because our model is dynamic, it can be used by a network to ascertain the sufficient threshold of minimal stock needed to attain a self-sustaining state. We consider several approaches to tip the network:

  – Initial User Content. We find 10.7% of the current observed levels of the ma-

ture UGC network is sufficient to tip the network. Marketing strategies oriented at inducing initial user posts, such as advertising or incentives and rewards for posting, are not uncommon approaches to encourage user content.

  - Initial Sponsored Content. An alternative option to tip the network is to substitute SSC for UGC; that is, sponsor content in an effort to attract posting and reading. Two strategies exist to achieve this outcome.
    * A firm can seek to jump start the network by sponsoring content early on, and then cease once the network self-sustains. Such an approach can minimize expenses as the cost of creating content is only borne by the firm early in the life of the network. We find that the initial period approach requires 9.3%.[3]
    * Alternatively, a firm can sponsor posts at a constant level, and thus change users' beliefs about steady state content. We find that the steady state strategy requires 8% the current observed levels of UGC to tip the network.

- Sponsored Content Quality. For both mature and nascent networks, site sponsored content can have higher average quality than UGC. In our counterfactuals, we consider the implications of higher quality posting strategies by the firm and find that tipping points can be substantially lowered. For the initial stock strategy, higher quality can lower the tipping point from 9.3% to 1% and for the steady stock strategy, the tipping point can be lowered from 8% to 1%. Higher quality posts can also be especially effective at growing mature networks.

Our model, owing to its dynamic structural nature and its ability to capture user expectations, is the first to our knowledge to shed light upon the role of different strategies in network tipping in the context of UGC. Overall, results suggest tipping is quite feasible with a relatively small amount of high quality sponsored content. This strategy may be quite cost effective as the expense in creating content manifests only in the early stages of the network. Examples of web sites that have pursued this strategy include `Soulrider.com` (Shriver et al.

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

2013), a wind surfing site which jump started the network by inviting experienced surfers to generate high quality content in its infancy. Of course, to grow the network, this site could have alternatively provided a smaller and more continuous stream of sponsored content, or instead targeted regular users with incentives to join the network, and our model yields insights regarding the potential of these various approaches to build the network.

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

In the next section, we elaborate upon how our model of user engagement differs from prior work on social networking in general, and user generated content in particular. We then discuss our data and context and use this information to construct our model. Then we explore some of the theoretical properties of our model, discuss identification and estimation, detail our results and conduct policy simulations regarding the effect of site sponsored content on reading and user generated content.

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

2 Literature Review

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

Likewise, our research is related to the burgeoning literature on user generated content (Albuquerque et al. 2010, Chevalier and Mayzlin 2006, Dellarocas 2006, Duan et al. 2008, Shriver et al. 2013, Ghose and Han 2011, Moe and Schweidel 2012, Zhang and Sarvary 2011, and Zhang et al. 2011) that considers the joint behavior of content consumption and generation.[5] Our research extends this work by developing a dynamic structural model of UGC; specifically, our model allows users beliefs about site engagement to evolve with the state of the network. This is material because changes in beliefs regarding the number of users contributing, for example, can affect whether agents visit a site, consume, or write. If interventions change these beliefs, it stands to reason that the behaviors of the agents will change. It is therefore necessary for any policy intervention to accommodate potential changes in beliefs, such as network tipping. Moreover, given that user generated content, like advertising, decays in efficacy over time and that those who post develop expectations about the likelihood their content is read in the future, there is considerable potential for dynamic behavior to be evidenced in the context of UGC.

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

As a dynamic structural model, our work is similar to Huang et al. (2011) who consider the blogging behavior of the employees of an IT firm. An important point of difference is that we use a rational expectations equilibrium framework to link individual behavior to aggregate state transitions (such as total posting and reading). In contrast, Huang et al. (2011) model users' behaviors independently of how others at the site react. In addition, our work extends the structural literature by considering how the quality of posts, in addition to the quantity, affect user engagement. As a result, we can conduct policy analyses on the role of site quality on user engagement.

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

Though we draw upon the approximate aggregation approach, our work fundamentally differs from past instantiations in many respects. First, a single unit of supply (posts) can be consumed by many (reads). 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 differing network effects, such as competition for readers (direct network effects) and the attraction of readers (indirect network effects). Second, because content is generated and consumed by a single agent, our problem differs considerably from previous markets wherein producers and suppliers differ. Third, we adapt the concept to an entirely new context, UGC networks.

Finally, because we consider strategies by which the UGC platforms can become self-sustaining, our work is related to the literature on tipping(Dubé et al. 2010;Katz and Shapiro

1994). In that research, tipping is defined as "the degree of market share concentration due to indirect network effects," Dubé et al. (2010) (p. 216). In our context, the indirect network effects for the platform arise from reading and posting rather than software and hardware. Our work is also structural in its foundation. In addition, we also consider user heterogeneity, which introduces a number of computational challenges that we solve using approximate aggregation.

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

3 Model

3.1 Model Overview

Figure 1 outlines the modeling context. Users consume content 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). Hence, we consider a model of information consumption predicated upon heterogeneous quality of UGC. We discuss and model this indirect network effect in Section 3.2.

An increase in content consumption can lead to an increase in content generation because those who post content presumably do so because they are motivated to have their posts read by others (Bughin 2007; Hennig-Thurau et al. 2004; Moe and Schweidel 2012; Nardi et al. 2004; Nov 2007) and we model this process in Section 3.3. There is also a potential direct network effect of content generation on content generation; as more content appears competition for readers increases; this problem is also considered in Section 3.3.

Last, it is theoretically possible for an increase in users to lead to more use via word of mouth (a direct effect of consumption on consumption). Our users are geographically dispersed and typically know each other only by their user ids on the site. Therefore they primarily contact other users on the site via an inter-mediating effect of posts, but this interaction occurs only through the content they generate and is thus an indirect network

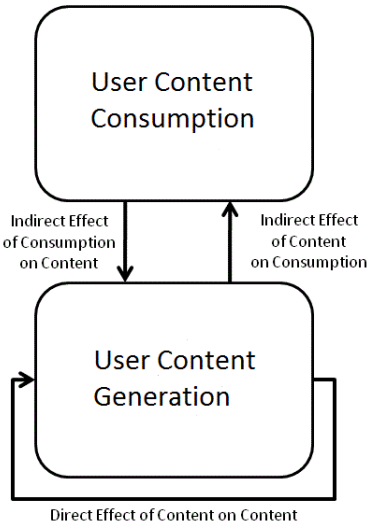

Figure 1: Model Overview

effect mediated by the total amount of UGC available. Moreover, given the content is freely available online and any piece of information can be read by multiple users at the same time, we expect no competitive effect between readers for content, thereby mitigating another source of potential direct network effects for consumption.

Both the decisions to generate and consume content are incumbent upon the decision to visit the site on a given day. To the extent the utility from visiting the site (stemming largely from the expected discounted utility of reading and posting there) exceeds that of outside options, users visit the site. We discuss this process in Section 3.4.

In sum, we consider i) M users' decisions (users are indexed by $i = 1, \ldots, M$) to visit a content sharing web site, n_{it}, on occasion t ($t = 1, \ldots, T$) and, ii) conditioned on that site visitation decision ($n_{it} = 1$), how much content to consume, r_{it}, and iii) how much content to generate, a_{it}. Users choose each of these three actions $\{n_{it}, r_{it}, a_{it}\}$ to maximize their utility conditioned on their beliefs regarding overall participation of others in the network. Though the decision to visit the site is made first, this decision is contingent on beliefs regarding the utility of reading and posting obtained after visiting. Hence, we solve this problem via backward induction and, in the ensuing discussion, first present the reading model in Section 3.2 and the posting model in Section 3.3, and then the visitation model in Section 3.4.

3.2 Content Consumption

3.2.1 Reading Utility

Readers consume content when the benefit of consumption exceeds its cost. The utility of reading is incumbent upon the total content available because an increase in the number of others' posts enhances the likelihood that a user finds items of interest. To formalize this notion, our model uses order statistics for post quality which follows 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, on this forum, ranging in quality from L (lower bound on content quality) to U (upper bound on content quality). The quality levels of the K_t posts are assumed to be uniformly distributed on a closed interval $[L, U]$, where $U > L \geq 0$ and $U - L$ is the support of the content quality distribution. As such, the qualities of the entire stock of postings, denoted as $Q_1, \ldots, Q_{K_t}$, are iid from $Uniform\,[L, U]$.[6]

Individuals read the content with the highest quality. Let the qualities of K_t content postings be ranked as their order statistics $Q_{[1]} \leq Q_{[2]} \leq \cdots \leq Q_{[K_t]}$ where $Q_{[1]}$ is the quality of the highest post and $Q_{[K_t]}$ is the quality of the lowest post. Each order statistic, $Q_{[k]}$, has the following distribution:

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

This order statistic is a linear transformation of the Beta distribution (note that $(Q_{[k]} - L)/(U - L)$ has a $Beta(k, K_t + 1 - k)$ distribution), so the expected quality for each order statistic is given by

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

Therefore, if individual i reads the r_{it} highest quality postings, the expected utility that reader obtains is

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

[6] In Section 7.1 we generalize our analysis to consider allowing site sponsored content to differ in its quality from user generated content.

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

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

The utility of reading can be further simplified when K_t is a large number using the approximation $K_t + 1 \approx K_t$ and $\left[K_t + \left(\alpha_1 - \alpha_2\right)/\alpha_1 + \alpha_2/2\alpha_1\right]/\left(K_t + 1\right) \approx 1$; this is the situation in our data because the total number of posts is large. Using this approximation results in reader i's utility for reading r_{it} posts following a quadratic formula

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

The reading utility is higher when α_1 which indicates the upper limit, U, on perceived content quality is higher and lower when $\alpha_2 = U - L$, which represents the uncertainty of the post quality, is higher. Given α_1 and α_2, the marginal utility of reading, which is the lowest quality post user i reads, is increasing with posts K_t. The result follows intuitively from a greater likelihood of finding content of interest simply when there are more posts. It is also interesting to note it provides a microeconomic foundation for empirical regularities detailed in recent research (Ransbotham et al. 2012). In sum, this utility evidences diminishing marginal returns from reading at a decreasing rate in the total number of posts and higher quality. Notice that we do not assume all the users have the same ordering of post quality.

3.2.2 Reading Costs

Next we consider the cost of reading. Following Yao and Mela (2008) and others, we assume the cost has a quadratic form that reflects an increasing scarcity of time or attention as more items are read:

$$c\left(r_{it}\right) = \left(\zeta_i + \kappa_{1it}\right) r_{it} + \kappa_{2i}\frac{r_{it}^2}{2}, \tag{6}$$

where κ_{1it} and κ_{2i} both being positive implies a convex cost function. We model heterogeneity in the reading cost function via a random effect for unobserved time-invariant heterogeneity ζ_i. Cyclicality, such as the weekend effect, is accommodated by allowing κ_{1it} in equation (6) to vary over time, that is $\kappa_{1it} = \kappa_{1i} w_t$ where w_t is a weekend indicator = 0 if weekdays, 1 if weekends. The cost parameters κ_{1i} and κ_{2i} are indexed by i because these can be heterogeneous across users. To model this heterogeneity in ζ_i, κ_{1i} and κ_{2i}, we assume there

are J segments of users and use a finite mixture model by letting these terms follow a discrete distribution,

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

Here user i belongs to class j with probability p_j. We constrain $\bar{\zeta}_j$ such that $\sum_{j=1}^{J} \bar{\zeta}_j = 0$ because α_1 and α_2 already subsume a non-zero mean for readings.

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

$$u\left(r_{it}\right) - c\left(r_{it}\right) = \left(\alpha_1 - \zeta_i - \kappa_{1it}\right) r_{it} - \left[\frac{\alpha_2}{K_t} + \kappa_{2i}\right] \frac{r_{it}^2}{2}, \tag{7}$$

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

$$r_{it}^* = \frac{\alpha_1 - \zeta_i - \kappa_{1it}}{\alpha_2 / K_t + \kappa_{2i}}. \tag{8}$$

Given heterogeneity in reading costs across segments, r_{it}^* differs across segments. After user i decides to visit the web site, she realized a contextual shock, ν_{it}, which is not observed by the econometrician. We assume the observed amount of reading by i is r_{it}^* multiplied by individual specific random shock (ν_{it}) so that $r_{it} = r_{it}^* \nu_{it}$.[7] Therefore, ex post amount of reading will also differ across different users in the same segment. As ν_{it} is realized after the user's site visitation 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 (8).

In sum, a reader's optimal level of consumption increases with the overall level of content on the site and the quality of posts.

3.3 Content Generation

3.3.1 The Per-Period Utility of UGC

Site users derive utility from others reading their posts and this expected posting utility is incumbent upon the users' beliefs their postings will be read. The expected amount of reading per posting 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 postings

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

(there is not a counter of "number of views" in our data).[8] This expected amount of reading per posting (y_t) is defined by

$$y_t = \frac{R_t}{K_t} = \frac{\sum_{i=1}^{M} E(n_{it} r_{it}^* | K_t, \zeta_i)}{K_t}. \tag{9}$$

Equation (9) demonstrates two competing effects of aggregate UGC K_t on y_t. First, there is a primary demand effect of K_t in the numerator as the expected optimal amount of reading increases with the supply of content, K_t. This constitutes an indirect effect of reading on posting.[9] Second, there is a competitive effect of K_t in the denominator as more postings will reduce the amount of reading per posting. 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 decay 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 geometric decay formulation is tantamount to an assumption that the probability of obsolescence for each post in each period is given by $1 - \rho$. The aggregate stock of posts therefore has the formulation of form:

$$K_t = \sum_{\tau=0}^{t} \rho^{t-\tau} A_\tau = \rho K_{t-1} + A_t \tag{10}$$

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

$$k_{it} = \sum_{\tau=0}^{t} \rho^{t-\tau} a_{i\tau} = \rho k_{i,t-1} + a_{it}. \tag{11}$$

[8] In Appendix A, we show that the users' expected amount of reading per posting y_t can be closely approximated by the exactly observed amount of reading per posting under the assumption of rational expectations, when the number of users and the UGC stock K_t are both very large - so actual reading rates can be used in our model estimation. However, it is necessary to recompute this rational expectation equilibrium in our counterfactual analyses (See Section 3.6 for more details). For other UGC websites, where there is a counter of number of views for every post, we can extend our model by let the expected amount of reading per posting be individual specific, y_{it}.

[9] Some agents might be inclined to post more if others are also posting. They have to see others' posts to know this, so the indirect effect of reading might also embed this tendency. We thank an anonymous reviewer for this insight.

Given users form rational expectations for the amount of reading per post y_t, the current period expected utility from generating content in period t can be written as the product of the number of the posts a user i writes and the rates at which these posts are read,

$$u(a_{it}) = g(k_{i,t}y_t) = g\left(\left[\rho k_{i,t-1} + a_{it}\right] y_t\right), \tag{12}$$

where $g(x)$ is a utility function with diminishing marginal return.[10] A common choice of $g(x)$ is

$$g(x) = \frac{x^{1-\gamma}}{1-\gamma}, \gamma \in [0, \infty) \tag{13}$$

where we have $g(x) = x$ when $\gamma = 0$ and $g(x) = \log(x)$ when $\gamma = 1$.

3.3.2 UGC Posting Costs, Heterogeneity, and cyclicality

The cost of writing is specified as

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

where the random error in the cost function, $\varepsilon_{it}(a_{it})$, has a generalized extreme value (GEV) distribution. The time-invariant component of the linear marginal cost ξ_i is allowed to be heterogeneous across users to accommodate users' different propensities to post and is assumed to follow a discrete distribution for a latent segment model. In addition, τ_{it} models cyclical effect such as a weekday effect, $\tau_{it} = \tau_i w_t$, where w_t is a weekday 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 (e.g., product involvement Dichter 1966 or anxiety Sundaram et al. 1998).

We integrate the heterogeneous parameters in the reading and posting into a joint discrete distribution with J latent segments,

$$[\zeta_i, \kappa_{1i}, \kappa_{2i}, \xi_i, \tau_i] \sim \sum_{j=1}^{J} p_j I\left(\zeta_i = \bar{\zeta}_j, \kappa_{1i} = \bar{\kappa}_{1j}, \kappa_{2i} = \bar{\kappa}_{2j}\right) I\left(\xi = \bar{\xi}_j, \tau_i = \bar{\tau}_j\right). \tag{15}$$

In the formula above, $\bar{\xi}_j$ is the segment-specific value of the time-invariant effect in the marginal cost of posting if user i belongs to segment j and $\bar{\tau}_j$ is the segment-specific cyclical

[10] Owing to the large number of observations at many sites such as the one we consider, sampling variation in y_t is very small given K_t. With smaller samples, it becomes necessary to integrate over uncertainty in y_t. That is, for large samples $y_t = R_t(K_t)/K_t = (\sum_{i=1}^{M} r_{it}^* \nu_{it} \to M r_{it}^*)/K_t$.

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

Finally, we define the non-random part of cost $c_{it}(a_{it})$ to be

$$\bar{c}_{it}(a_{it}) = \tau_{it} + \xi_i a_{it}.$$

3.3.3 Optimal UGC Posting

As past and current postings can be read by others in the future, a user's posting decision is inherently dynamic. As such, we presume a user chooses the number of postings a_{it} (amount of content to generate) that maximizes the discounted expected sum of per-period utilities minus per-period costs to obtain the following value function

$$V_i(s_{it}, \varepsilon_{it}) = \max_{a_{it}, a_{i,t+1}, \ldots} E\left\{\sum_{k=t}^{\infty} \left[u(a_{ik}) - c_{ik}(a_{ik})\right]\right\}. \tag{16}$$

In this dynamic optimization problem, $s_{it} = \{k_{i,t-1}, K_t, \tau_{it}\}$ and ε_{it} are the state variables and the number of per-period postings a_{it} is the control variable.[11] Posting by users a_{it} is treated as a discrete variable because over 99% of users post only from zero to ten posts a day (see Section 4 for more detail on the data). Let $A \equiv \{0, 1, \ldots, \bar{a}\}$ be the action space where $\bar{a}$ is a large integer representing the upper bound of postings a user can write in period t. Therefore the posting decision a_{it} is a discrete choice of number of postings, i.e., $a_{it} \in A = \{0, 1, 2, \ldots, \bar{a}\}$,

[11] The inter-temporal substitution of posting for the dynamic optimization problem is as follows. If we treat a_{it} as a continuous variable, we can derive the Euler equation $-\left[\left(y_t\left(\rho k_{i,t-1} + a_{it}\right)\right)^{-\gamma} - \left(\tau_{it} + \xi_i\right)\right] / (\tau_{i,t+1} + \xi_i) = \beta\rho$, which shows posting one more unit of UGC will gain utility $\left[y_t\left(\rho k_{i,t-1} + a_{it}\right)\right]^{-\gamma}$ and incur cost $(\tau_t + \xi_i)$ in the current period t. This additional posting will also gain utility discounted by $\beta\rho$ in the next period $t+1$. However, if a user selects to post in $t+1$ instead of t, she will forgo the utility in t, which is $\left[y_t\left(\rho k_{i,t-1} + a_{it}\right)\right]^{-\gamma}$, and avoid cost $(\tau_{it} + \xi_i)$. The cost incurred in $t+1$ is $(\tau_{i,t+1} + \xi_i)$ instead. The optimal number of postings is achieved when the user is indifferent about whether posting an additional unit in t or $t+1$. Thus, increasing durability of UGC, ρ, tends to increase the incentive to post in the current period. However, the competitive effect from the increased postings of other users and one's own past postings constitutes indirect disincentive to post.

The value function of this optimization problem in the form of Bellman's equation is

$$V_i(s_{it}, \varepsilon_{it}) = \max_{a_{it} \in A} \{u(a_{it}) - \bar{c}_{it}(a_{it}) + \varepsilon_{it}(a_{it}) + \beta E[V_i(s_{i,t+1}, \varepsilon_{i,t+1}) | s_{it}, a_{it}]\}. \quad (17)$$

where $E[V_i(s_{i,t+1}, \varepsilon_{i,t+1})|s_{it}, a_{it}]$ represents the expected future value given the current states and $u(a_{it}) - \bar{c}_{it}(a_{it}) + \varepsilon_{it}(a_{it})$ represents the current period net utility of action a_{it}. The solution to this Bellman equation results in an individual's optimal posting policy, denoted by $\sigma_a^*(s_{it}, \varepsilon_{it})$, and reflects how posting levels change with the amount of aggregate posting stock and an individual's own posting stock.

3.3.4 Posting Level Probabilities

To derive the probability of observing a user posting a specific number of posts under this optimal decision rule, we first define the integrated value function $\tilde{EV}_i(s_{it}, a_{it})$ as

$$\tilde{EV}_i(s_{it}, a_{it}) = \int_{s_{i,t+1}} \int_{\varepsilon_{i,t+1}} V_i(s_{i,t+1}, \varepsilon_{i,t+1}) p(s_{i,t+1}, \varepsilon_{i,t+1} | s_{it}, \varepsilon_{it}, a_{it}) ds_{i,t+1} d\varepsilon_{i,t+1}. \quad (18)$$

The $\tilde{EV}_i(s_{it}, a_{it})$ function represents the future value of a given user's action under a set of given states. Using the conditional independence assumption for ε_{it} and s_{it}, this function can be simplified as follows,

$$\tilde{EV}_i(s_{it}, a_{it}) = \quad (19)$$
$$\int_{s_{i,t+1}} \log \left\{ \sum_{a_{i,t+1} \in A} \exp \left[u(a_{i,t+1}) - \bar{c}_{i,t+1}(a_{i,t+1}) + \beta \tilde{EV}_i \left(s_{i,t+1}, a_{i,t+1} \right) \right] \right\} p(s_{i,t+1} | s_{it}, a_{it}) ds_{i,t+1}.$$

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

$$u(a_{it}) - \bar{c}_{it}(a_{it}) + \varepsilon_{it}(a_{it}) + \beta \tilde{EV}_i(s_{it}, a_{it}) \geq$$
$$u(a'_{it}) - \bar{c}_{it}(a'_{it}) + \varepsilon_{it}(a'_{it}) + \beta \tilde{EV}_i(s_{it}, a'_{it}), \forall a'_{it} \neq a_{it} \in A, \quad (20)$$

by which we derive the probability of writing a_{it} content postings conditional on site visitation as[12]

$$P(a_{it}|s_{it}, n_{it} = 1) = \frac{\exp(u(a_{it}) - \bar{c}_{it}(a_{it}) + \beta \tilde{EV}_i(s_{it}, a_{it}))}{\sum_{a'_{it} \in A} \exp(u(a'_{it}) - \bar{c}_{it}(a'_{it}) + \beta \tilde{EV}_i(s_{it}, a'_{it}))}. \tag{21}$$

3.4 Site Visitation

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

$$\begin{aligned} u(n_{it} = 1) &= \mu_1 E \max_{r_{it}} [u(r_{it}) - c_{it}(r_{it})] + \\ &\quad \max_{a_{it}} [u(a_{it}) - c_{it}(a_{it}) + \beta \tilde{EV}_i(s_{it}, a_{it})] + \eta \varepsilon_{it}(n_{it} = 1) \end{aligned} \tag{22}$$

where μ_1 is a scale parameter that rescales the utility of reading relative to the utility of posting.[13] 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 using the site ($n_{it} = 0$) contains three components. First, users continue to obtain utility from those who read their content generated from past visits. The utility from others' reading previous post stock is given by $g(k_{it}y_t) = g(\rho k_{it-1}y_t)$ with the attendant reading rate y_t in period t. Second, μ_{0i} is a segment-specific intercept which 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 of not visiting the site is obtained by summing these three components:

$$u(n_{it} = 0) = \mu_{0i} + g(k_{it}y_t) + \beta \tilde{EV}_i(s_{it}, n_{it} = 0) + \eta \varepsilon_{it}(n_{it} = 0), \tag{23}$$

where $\tilde{EV}_i(s_{it}, n_{it} = 0)$ is also an integrated value function. A user chooses to visit the web

[12] Because of the cost function in (21), contiguous postings will have more similar choice probabilities than distant ones and are conditionally dependent. To see this, note that $\Delta\bar{c} = \bar{c}(a) - \bar{c}(a') = \xi(a - a') - (\varepsilon(a) - \varepsilon(a'))$. As the last term is zero in expectation, we can see that $\Delta\bar{c}$ is strictly increasing in the distance between two levels of posting.

[13] Regarding the expected utility of reading, note that equation (8) assumes that users who visit a site will always read at least some posts, because $r^*_{it} > 0$ is always an interior optimal solution. This implies the expected utility of reading is always greater than zero if a user decide to visit. This specification is consistent with the data as 99.998% of the users read postings upon entering the site.

site if $u(n_{it}=1) > u(n_{it}=0)$ and vice versa. This equation yields an optimal decision rule mapping the choice of whether to visit the site to observed states, $\sigma_n^*(s_{it}, \varepsilon_{it})$.

To derive the choice probabilities for site visitation, we assume $\varepsilon_{it}(a_{it})$, $\varepsilon_{it}(n_{it}=0)$ and $\varepsilon_{it}(n_{it}=1)$ have iid Type-1 Extreme Value (Gumbel) distributions, resulting in a nested logit model of site visitation and content generation given site visitation.[14]

Because not visiting the site is equivalent to visiting the site and having zero postings, we have

$$\tilde{EV}_i(s_{it}, n_{it}=0) = \tilde{EV}_i(s_{it}, n_{it}=1, a_{it}=0). \tag{24}$$

We define the inclusive value of writing content postings conditional on site visitation as

$$IV_{it} = \ln \sum_{a_{it} \in A} \exp(u(a_{it}) - \bar{c}_{it}(a_{it}) + \beta \tilde{EV}_i(s_{it}, a_{it})). \tag{25}$$

Based on equations (22), (24), and (25), we derive the choice probability of visiting the site as

$$P(n_{it}=1|s_{it}) = \frac{\exp\{\mu_1 E \max_{r_{it}}[u(r_{it}) - c_{it}(r_{it})] + \eta IV_{it}\}}{\exp\left\{\mu_0 + \eta\left[g(k_{it}y_t) + \beta \tilde{EV}_i(s_{it}, n_{it}=0)\right]\right\} + \exp\{\mu_1 E \max_{r_{it}}[u(r_{it}) - c_{it}(r_{it})] + \eta IV_{it}\}} \tag{26}$$

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

Note that when we apply the latent segment model in equation (15), the integrated value function $\tilde{EV}_i(s,a)$ is the same for all the users in segment j ($j = 1, \ldots, J$). Hence, we let $\tilde{EV}_i(s,a) = \tilde{EV}_j(s,a)$ if user i is in segment j.

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 the day of week effects, w_t evolves deterministically with the time. Lastly, the aggregate stock of UGC, K_t, is defined as $K_t = \sum_{i=1}^{M} k_{it}$ and hence it evolves deterministically given K_{t-1} and

[14] An alternative model for the random errors assuming McFadden's Generalized Extreme Value (GEV) distribution for $\varepsilon_{it}(a_{it})$ and $\varepsilon_{it}(n_{it}=0)$ and $\varepsilon_{it}(n_{it}=1)=0$ yields an equivalent nested logit model with the inclusive value function and choice probabilities subject to reparameterization. See Choi and Moon (1997) for the details of the inclusive value function for the GEV model.

a_{it}; $i = 1, \ldots, M$. However, from the perspective of any individual user i, K_t given K_{t-1} is stochastic because she does not observe other users' a_{it} and ε_{it}. When the site has a very large number of users, each user i believes her own action a_{it} has no influence on the aggregate UGC, K_t. This claim is similar to the assumption of pure competition where no agents in the market assume their individual output can change the total supply. Hence, user i assumes that K_t evolves somewhat stochastically given K_{t-1}, but independent of their own action a_{it}. If we impose a rational expectations constraint, then user i's belief about the state transition for K_t must coincide with the actual behavior by users on the site. This will be discussed in detail next.

3.6 Rational Expectations Equilibrium and Approximate Aggregation

Rational expectations require that users' beliefs about K_t be consistent with its actual transition, which reflects the sum of all individuals' posting decisions. This observation becomes critically important in policy simulations because there is no reason to presume the evolution of K_t is invariant to a change in policy that might affect users' participation levels; that is, 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

$$K_t = \omega_0^K + \omega_1^K K_{t-1} + \omega_2^K w_t + \varepsilon_t^K, \tag{27}$$

where w_t is the weekend indicator and ω_2^K represents cyclical effect. The parameters $\omega_0^K,\ \omega_1^K, \omega_2^K$ 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 help ensure that the approximate beliefs regarding the aggregate state transitions are consistent with the Markovian structure in the underlying

individual posting model.[15] 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 individual users approximate the expected average amount of reading per posting as a function of K_t with equation

$$y_t = \omega_0^y + \omega_1^y K_t + \omega_2^y w_t, \tag{28}$$

where ω_2^y is again for the cyclical effect of weekend. Equation (28) approximates equation (9) 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 posting can closely approximates the expected one. See Appendix A for details. The parameters $\omega_0^y,\ \omega_1^y, \omega_2^y$ are also determined by the rational expectations equilibrium.

3.6.2 Exact Aggregation

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

$$K_t = \rho K_{t-1} + \sum_{i=1}^{M} a_i \left(k_{it}, K_t, \tau_{it}, \varepsilon_{it}\right). \tag{29}$$

Using equation (29) directly to compute users' rational expectations requires us to assume all users know all other users' policy functions $a_i \left(k_{it}, K_t, \tau_{it}, \varepsilon_{it}\right)$ as well as the distribution of their individual-level posting stock k_{it}. Complete knowledge of the behavior of thousands of other users is an unrealistic assumption, which imposes a large informational burden on every individual user. In addition, this assumption places the distribution of k_{it} in every user's set of state variables. Because the distribution of k_{it} is high-dimensional, the "curse of dimensionality" renders the dynamic programming problem intractable. On the other hand, approximate aggregation (assuming bounded rationality) only requires that K_t and y_t

[15]Note that the order of the state transition equations cannot be higher than the order of the individual level model, else the individual level model would fail to account for consumer's beliefs about these higher order states. Here we assume individuals only use one lagged K_{t-1} to predict K_t and hence it implies an AR(1) model for K_t. Conceivably, individuals may use more than one lagged stock to predict K_t.Were they to use an AR(q) model then $K_{t-2}, \ldots, K_{t-q}$ would also have to be in the set of state variables in the dynamic optimization problem. As a surfeit of state variables can induce computational dimensionality constraints, the most parsimonious state transition model possible for K_t is desirable from a computational perspective. In Section 4.2, we test and find the AR(1) model is best model for our data.

computed from equations (28) and (29) respectively in the individual optimization problem coincide with K_t and y_t computed from the exact aggregation. This implies agents need only be able to form rational beliefs regarding the transitions of the aggregate states. Krusell and Smith (1998) show that using the state transition rule such as in equation (29) can still generate a stationary distribution of k_{it} instead of a degenerate k_t for every agent (which we further confirm by simulation in Appendix D.2.1).

3.6.3 Consistency Between Exact and Approximate Aggregation

The approximate aggregation approach requires that we ensure that the beliefs regarding aggregate state transitions are consistent with the individual behaviors that underpin it. Using an initial guess for the parameters $\omega_0^K,\ \omega_1^K, \omega_2^K$ and $\omega_0^y,\ \omega_1^y, \omega_2^y$, we compute individual behaviors n_{it}, a_{it} and r_{it}^*. Aggregating across persons, we recompute K_t and y_t and recompute individual behaviors, iterating back and forth between the individual-level and aggregate models until convergence. Appendix B details the algorithm used to compute a rational expectations equilibrium. The parameters $\omega_0^K,\ \omega_1^K, \omega_2^K$ and $\omega_0^y,\ \omega_1^y, \omega_2^y$ are re-estimated in every step of the iterations to find the fixed point of the rational expectations equilibrium.[16] In sum, the use of approximate aggregation enables us to accommodate heterogeneity in a rational expectations equilibrium model.

3.7 Network Effect of Aggregate UGC

Thanks to its structural foundation and inclusion of direct and indirect network effects, our model evidences flexibility in characterizing aggregate UGC behaviors. We exemplify this point below by considering i) the marginal effect of additional content and ii) the role of initial stock. We explore other theoretical implications of our model in Appendix D.[17]

A key consideration in assessing the role of UGC on overall site traffic is the marginal effect of additional content on consumption. The indirect network effect of the aggregate

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

[17] The appendix details i) convergence to the defined rational expectations equilibrium in Section 3.6 and ii) how the model's parameters influence the network's user content and readings in equilibrium, including the role of initial content on network size and the effect of content stock decay on content generation.

UGC on an individual user's posting action is affected by the likelihood her post is read; that is, the numerator (aggregate reading) and denominator (aggregate postings) in equation (9). The numerator implies a greater likelihood of reading because more content, K_t, enhances the consumption experience. The denominator implies a competitive effect of K_t as content increasingly competes for users. As we show below, this indirect effect can be either positive or negative.

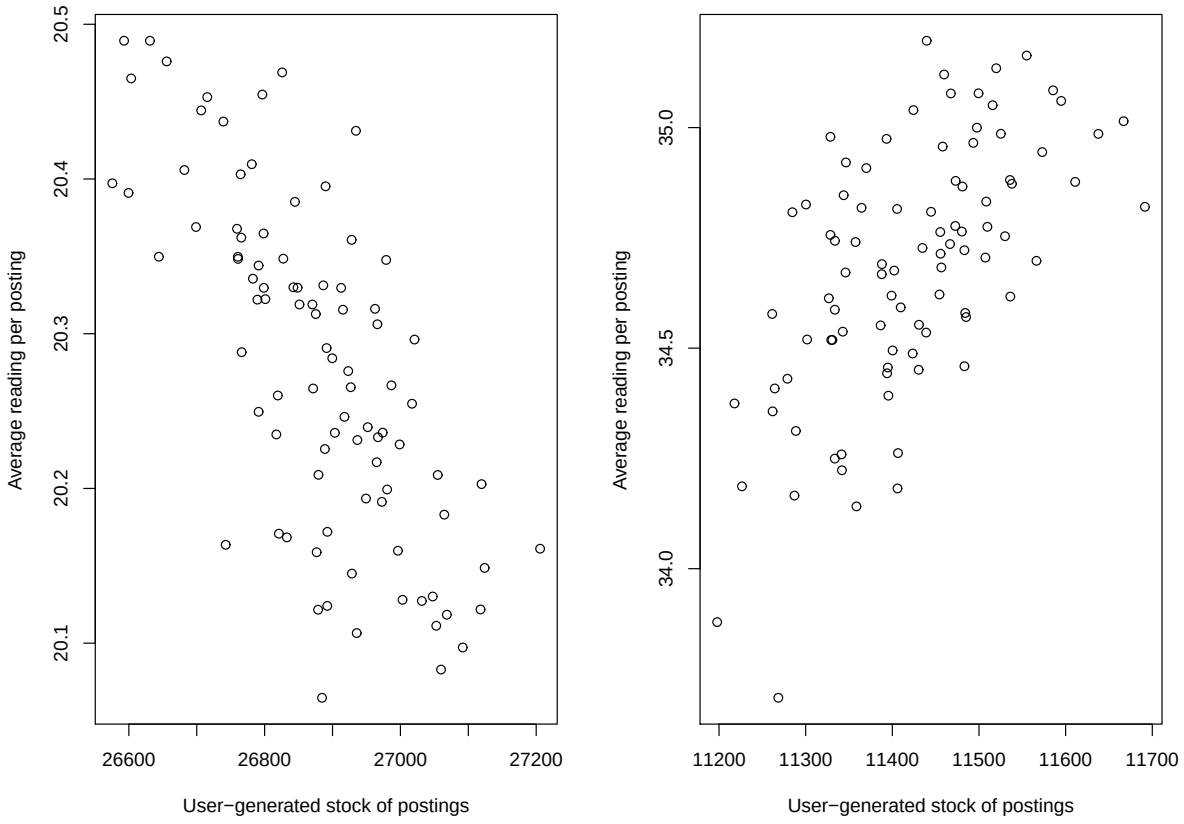

Figure 2: The relationship between average reading per posting and aggregate user-generated posting stock in equilibrium.

In Figure 2, we show two simulation examples, one where y_t is decreasing in K_t and another where it is decreasing. The decay rate ρ is 0.6 for the first example and 0.1 for the second: all the remaining parameter values are identical in the two examples.[18] We also find

[18]This simulation considers two segments in our 100 period simulation, each of whom have the same cost of

the relationship between y_t and K_t can switch sign if we adjust the ratio of population sizes of the two segments. Because the numerator is not a closed-form function of K_t, it is not clear under what conditions the network effect of K_t is positive. We conjecture that the positive indirect effect is more likely when there is a strong primary effect on site participation and that the negative indirect effect is more likely when the participation is already high.

Of note, a descriptive model of network effects (common in prior research) would lead to vastly different interpretations of the indirect network effect under these two scenarios. In contrast, the model we develop is sufficiently flexible to accommodate the change in the signs of these marginal effects, depending on the state of the network. The sign of the network effect at the current state of the network is especially important for the forum managing firms when they consider whether using sponsored content can increase traffic and UGC on their web sites.

4 Data

Our data come from a large Internet site devoted to a common interest, which includes a forum where persons can discuss various topics much like fans would discuss a sports team, its players or various games. 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 inclusive of zeros. This yields 19,461,572 user-day observations.

4.1 Descriptive Statistics

Table 1 reports the descriptive statistics for the key variables (excluding 0.05% of outliers) used in our analysis. The table indicates visitation is frequent, with 42% of users visiting the site on any given day. Forum reading is far more prevalent than forum posting, and there is significant variation in forum reading and posting across individuals as indicated by large standard deviations of these variables. The average number of individual posting stock

reading but vary in their posting costs and size (Segment 1 is smaller and has lower posting costs, consistent with the notion that a small number of users predominate the number of posts). Appendix D.1 details the specific parameters of our simulation.

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

Table 1: Descriptive Statistics

is quite low (1.19), but some users are heavily invested in the site with larger cumulative posting stocks.

Further considering the differences across users, we present the distribution of site engagement, defined as visitation rates, reading rates and posting rates. From Figure 3, we note the observed rates are remarkably close to the endemic "80/20" rule observed in many marketing contexts, 20% of the users are responsible for 76% of postings and 73% of readings. This observation again suggests the need to accommodate unobserved heterogeneity in reading and posting.

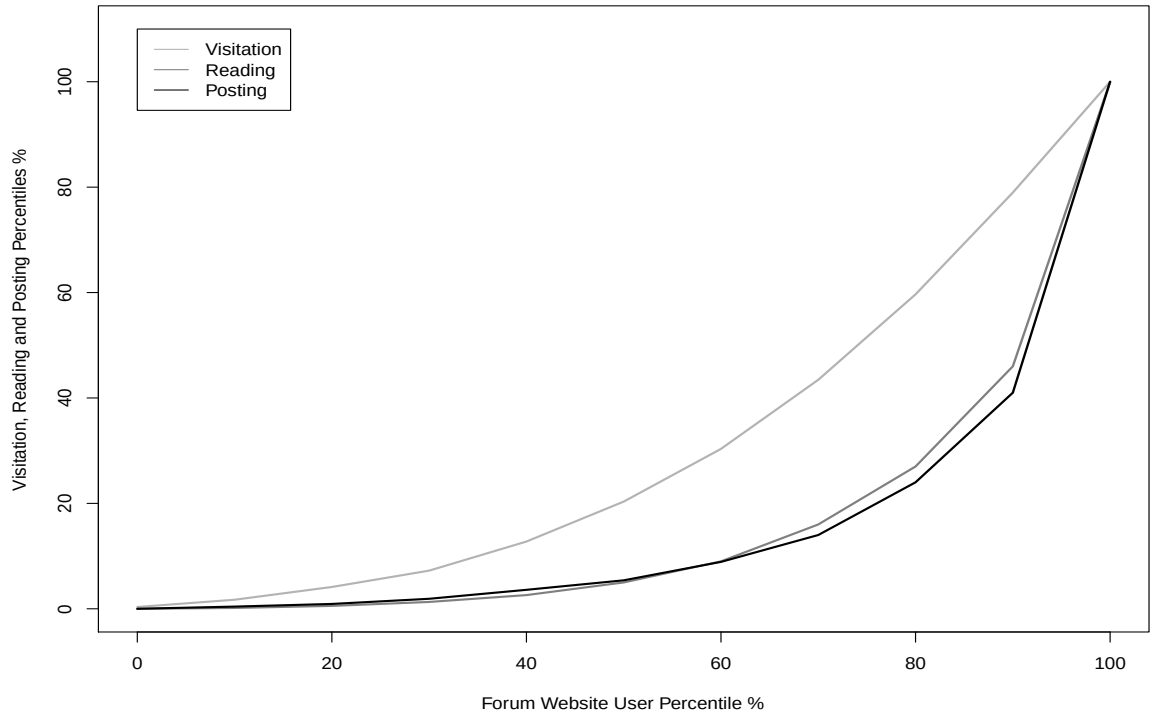

Figure 3: Percentile Plot of Log-in, Reading and Posting

Finally, Figure 4 plots the joint distribution of content generation and consumption, conditioned on site visitation. The figure indicates 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 6 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 visitation, all users read posts, but some do not create posts. Therefore, our reading model has a non-zero interior solution for the utility optimization problem, whereas we apply the discrete choice model with the option of choosing zero posts for the posting model.

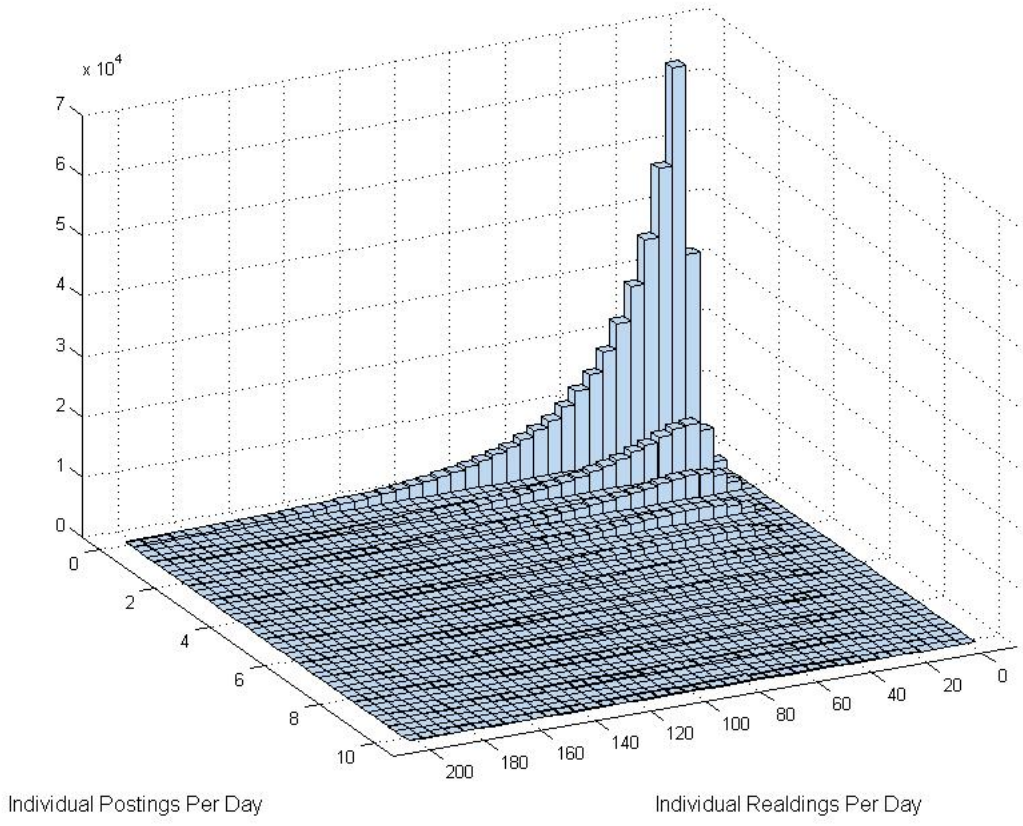

Figure 4: Joint Distribution of Reading and Posting

4.2 Exploratory Analysis

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

4.2.1 Reading

First, we consider reading. Recall, our model in Section 3.2 posited a positive link between aggregate content stock K_t and individual reading r_{it}. Hence, we regress the daily amount of reading of individuals against aggregate user-generated content stock K_t (unit in 10,000 postings) and individual post stock k_{it}, using a random data sample of activities of 600 users over 61 days. Table 2 demonstrateshigher aggregate user-generated content stock indeed leads to higher daily individual among of reading, whereas indivial post stock k_{it} does not significantly affect daily reading, consistent with the reading utility model.

Variable	Parameter Estimate	t-value	p-value
Aggregate UGC Stock	0.79^*	3.72	0.00
Individual Post Stock	0.11	1.32	0.19
Weekend Effect	-1.03^*	-2.03	0.04

Table 2: The Effect of Aggregate Posting Stock on Individual Reading

4.2.2 Posting

Second, we explore the effects of the average amount of reading per posting y_t, competitive effects of aggregate posting stock K_t and the individual posting stock on the individual postings as discussed in Section 3.7. Because daily individual postings are small integers, we fit a generalized linear model using the Poisson family. Note that the reading utility function suggests higher reading rates should increase the utility of posting; thus a finding that higher reading rates correlate with posting is consistent with our theory. In addition, since higher individual posting stock decreases the marginal utility of posting, we expect to see negative correlation between individual stocks and posts.

Table 3 reports the results of this regression. Consistent with our assumptions in Section 3.3, the results suggest a significant positive effect of average amount of reading per posting and a significant negative effect of individual posting stock on the likelihood of posting.

The aggregate UGC stock does not have a direct significant effect on individual posting conditioned on the average reading per posting, consistent with our specification of posting utility wherein posting is mediated by reading. Overall, this exploratory regression analysis is consistent with the content generation model.

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

Table 3: The Effect of Reading per Posting on Individual Posting

5 Estimation and Identification

5.1 Estimation

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

To alleviate 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 (27) and the reading-per-posting as a function of the UGC in (28), and second estimating the parameters in reading, posting and site-visitation models.

In the first step of this approach, we estimate the state transition equation for the aggregate UGC in equation (27) and the reading-per-posting as a function of the UGC in equation (28). We obtain the MLE estimates of the regression coefficients $\omega_0^K,\ \omega_1^K, \omega_2^K$ and $\omega_0^y, \omega_1^y, \omega_2^y$, which capture the evolution of the aggregate states in the data under the current equilibrium. Because the observed K_t and y_t reflect the current equilibrium, no iteration is needed to re-estimate $\omega_0^K,\ \omega_1^K, \omega_2^K$ and $\omega_0^y, \omega_1^y, \omega_2^y$. This equilibrium assumption is not likely to hold in our counterfactuals wherein we do need to recompute the expectations.

In the second step, we estimate the structural parameters in the individual reading, posting and site-visitation models (See Appendix (C.2) for details). In this step, we use the first step's empirical estimates of $\omega_0^K,\ \omega_1^K, \omega_2^K$ and $\omega_0^y, \omega_1^y, \omega_2^y$ in the state transition equation as indicated in equation (19) when estimating the structural parameters. The reading and posting models are estimated jointly with MLE, using the joint likelihood function of reading and posting for each individual user i

$$\left\{ \sum_{j=1}^{J} p_j \prod_{t=1}^{T} \text{Exponential}\left(r_{it} | \frac{\alpha_1 - \bar{\zeta}_j - \bar{\kappa}_{1j} w_t}{\alpha_2 / K_t + \bar{\kappa}_{2j}} \right) \frac{\exp\left(u(a_{it}|s_{it}) - \bar{c}(a_{it}|s_{it}) + \tilde{EV}_j(s_{it}, a_{it}) \right)}{\sum_{a'_{it} \in A} \exp\left(u(a'_{it}|s_{it}) - \bar{c}(a'_{it}|s_{it}) + \tilde{EV}_j(s_{it}, a'_{it}) \right)} \right\}. \tag{30}$$

The site-visitation model in equation (26) is estimated as a binary choice model with the likelihood function

$$\left\{ \sum_{j=1}^{J} p_j \prod_{t=1}^{T} \left[P(n_{it} = 1|s_{it})^{n_{it}} P(n_{it} = 0|s_{it})^{1-n_{it}} \right] \right\}, \tag{31}$$

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

To alleviate 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. (2009), which is a maximum likelihood estimator using mathematical programming with equilibrium constraints (MPEC). See Appendix C.2 for details. Su and Judd (2010) show that the two-step pseudo maximum likelihood (2S-PML) estimator discussed in the Appendix C is consistent (see Page 33 in Su and Judd 2010). We bootstrap to compute the 95% confidence intervals.

5.2 Identification

5.2.1 Posting Stock Decay Parameter ρ

As indicated in Section 3.3, the information contained in a posting gradually becomes obsolete. We model this phenomenon by imparting an exogenous decay parameter ρ to the posting stock in our model. The decay rate ρ is identified and estimated using a secondary data set, which records a sample of postings 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 amount per posting, aggregate posting stock and visitation rate are stationary in time. The data are consistent with our modeling assumption that a posting has a finite lifetime with a decay parameter ρ. When the average reading amount per posting is stationary over time as observed in our data, the decay parameter is identified by the ratio of the times that a posting 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 posting, i.e, the ratio of reading per posting in period t divided by reading per posting in $t+1$.[19]

5.2.2 Other Parameters

In the aggregate state transition equations (27) and (28), the coefficients are identified form the time series structure of the aggregate level data – specifically, from the autocorrelation for the K_t and correlation between the y_t and K_t. Next, we consider the identification of the parameters in the reading and posting models from the second stage estimation. Note that the optimal individual level reading level r_{it}^* for person i at time t is equal to $(\alpha_1 - \bar{\zeta}_j - \bar{\kappa}_{1j} w_t)/(\alpha_2/K_t + \bar{\kappa}_{2j})$ if person i belongs segment j per equations (8) and (15). This expression suggest 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 10, which also achieves scale normalization. Heterogeneity in the reading ($\bar{\zeta}_j$) 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 classify the days of a week into weekdays and weekends and use a dummy variable that is set 0 for weekdays and 1 for weekends. Hence, these parameters are identified by differences in the mean amount of reading and number of postings between weekdays and weekends. The parameter that captures the diminishing marginal returns for the utility in posting, γ, is identified by the observed difference in mean posting levels at

[19] 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 number of postings and the average amount of reading per posting at period t. For our data (see the results in Section 6.1), we find neither factor to be statistically significant. This suggests that the decay parameter ρ is independent of the aggregate activities of the forum and therefore exogenous.

different levels of posting stock. In general the discount factor is not identified in dynamic discrete choice models (Rust 1987, 1994). Hence we set $\beta = 0.99$.[20] Finally, there are scale parameters to estimate in the site visitation model μ_0, μ_1 and η. Conditioned 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 visitation.

6 Results

6.1 Estimation of Decay Parameter and Initialization of Posting Stock

As indicated in Section 3.3, the posting stock is incumbent upon the decay rate of a post. The exogenous posting decay parameter ρ is estimated using an auxiliary dataset collected by the Internet site regarding when a sample of users' posts were visited by other users. The decay in the number of users clicking on these posts over time is informative about their durability. From these data, we consider a random sample of 474 postings posted on the forums in the first week of sampling period.

The decay parameter is identified by the ratio of the times that a posting is read in periods $t-1$ and t. Under the constant decay assumption, this ratio equals the decay rate in the amount of reading per posting (the ratio of reading per posting in period $t-1$ divided by reading per posting in t) when the network is in a steady state. We also test the exogeneity of the decay parameter by regressing on the aggregate number of postings K_t and the average amount of reading per posting y_t at period t. Let z_{kt} be the number of times posting k is read is in period t, we estimate ρ using the following regression model by the feasible generalized least square (because $z_{kt}/z_{k,t+1}$ has larger variance when z_{kt} and $z_{k,t+1}$ are small) specified as follows:

$$\log \frac{z_{kt}}{z_{k,t-1}} = \log \rho + \beta_1 K_t + \beta_2 y_t + \beta_3 w_t + \varepsilon_t^z, \tag{32}$$

where w_t is the weekend indicator. The resulting estimate is $\widehat{\log \rho} = -0.31$ (sd = 0.0025), so the decay ρ is 0.74, which implies that 90% of the post's stock is depleted after one week. The estimate of the regression coefficient β_1 for the aggregate postings at period t is

[20] We also estimated the model with $\beta = 0.98$ and our findings and insights remain unchanged.

-3.06×10^{-6} with the p-value equal 0.85. The estimate of the regression coefficient β_2 for the average reading per posting y_t at period t is 0.062 with the p-value equal 0.56. Hence, neither of the factors are statistically significant. This allows us to confirm the hypothesis that the decay parameter is exogenous and independent of the aggregate activities of the forum when it is in steady state.

Note that individuals' initial posting stocks in the first week of the data are unobserved, as there is no history of posts prior to the initial week. Hence, using this posting stock decay estimate, the individual posting stock is computed by setting the initial stock at zero and recursively applying equation (11) using the 61-day posting data repeatedly until the individual's posting stock reaches a steady state. The individual's steady state is then re-used as the initial posting stock to calculate the individual posting stock for the 61-day data. We adopt this practice because most of the users in our sample have been using the forum for a long time prior to the sampling period, hence their posting stocks are likely to have reach the steady state (with daily random variation) at the inception of our data. Aggregate stock K_t is computed by aggregating individual stocks.

6.2 Results for 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 (27) and (28) are reported in Table 4. The results provide evidence of strong auto-correlation ($\omega_1^K = 0.93$) for the aggregate stock. The rate of reading per posting is an increasing function of aggregate stock ($\omega_1^y = 5.43 \times 10^{-6}$ is statistically significant), which implies the positive indirect network effect of posting on site participation exceeds the negative competitive effect for our forum data. The weekday effects for Monday, Tuesday and Wednesday in model (27) are not significantly different from Sunday, whereas the effects for Thursday, Friday and Saturday are significantly negative, which implies lower posting activity for these days of a week. Because these day effects differed only between the weekday and the weekend, we grouped them into a single weekend indicator for Thursday through Saturday. The weekend effect in model (28) is significantly negative, which means lower reading activity for these days.

We also test an AR(2) model $K_t = \omega_0^K + \omega_1^K K_{t-1} + \omega_2^K K_{t-2} + \omega_3^K w_t + \varepsilon_t^K$ using the second

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

Table 4: Estimation Results for Aggregate Posting Stock Transition Equation and Rate of Reading-per-Posting Equation (with 95% confidence intervals in brackets)

order lag K_{t-2}. We find the second lag coefficient ω_2^K to be not significant statistically (p-value $= 0.73$). Durbin-Watson test for the residuals of the AR(1) model $K_t = \omega_0^K + \omega_1^K K_{t-1} + \omega_2^K w_t + \varepsilon_t^K$ has the p-value equal to 0.63, which cannot reject the null hypothesis that the autocorrelation of the residuals is 0. Therefore, our data supports the assumption that users can use approximate aggregation and AR(1) to predict K_t on the basis of rational expectations.

6.3 Individual-level Model Results (Posting, Reading and Site Visitation)

We randomly select a sample of 600 users to estimate the individual-level model. The amount of reading and number of postings for each individual in the sample are recorded for 61 days from October 1st to November 30th, 2009. If both reading and posting are zero for a user in a certain day, we conclude the user does not visit the site that day. Table 5 reports the parameter estimates for the model assuming two segments of users.[21]

We begin by discussing the results from the posting and reading models The two segments of the posting 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. Likewise, the two segments of the reading model share a common utility parameter, α_1, but differ with respect to linear marginal cost parameter, $\bar{\zeta}_j$.

Comparing the two groups, the second segment is larger in size and evidences higher reading and posting costs; hence, this group of users read less often and rarely posts content.

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

Parameters	First Segment Frequent Users	Second Segment Light Users
Posting Model		
Utility coefficient γ	0.79 $[0.74, 0.85]$	
Cost coefficient $\bar{\xi}_j$	1.29 $[1.13, 1.47]$	6.83 $[6.02, 7.12]$
Weekend effect $\bar{\tau}_j$	8.99 $[1.12, 14.72]$	9.60 $[2.46, 14.75]$
Reading Model		
Utility coefficient α_2	0.60 $[0.48, 0.75]$	
Linear cost coefficient $\bar{\zeta}_j$	-6.96 $[-5.38, -7.62]$	6.96 $[5.38, 7.62]$
Weekend cost effect $\bar{\kappa}_{1j}$	$-0.50[-1.08, 0.01]$	-0.05 $[-0.39, 0.075]$
Quadratic cost coefficient $\bar{\kappa}_{2j}$	$0.26[0.21, 0.28]$	0.11 $[0.08, 0.15]$
Site Visitation Model		
Intercept, μ_{0j}	$27.12[1.01, 53.94]$	4.46 $[0.37, 11.22]$
Reading scale parameter μ_1	0.051 $[0.0031, 0.13]$	
Gumbel scale parameter η	0.97 $[0.94, 0.99]$	
Heterogeneity		
Segment size	43% [38%, 48%]	57% [52%, 62%]

Table 5: Estimation Results for Utility and Cost Parameters in Posting and Reading

Hence, we denote them "light users." Also of note, the weekend effect $\bar{\tau}_j$ in the posting cost function is positive, so the users tend to post less on a weekend.

Next, we consider the parameter estimates in the site visitation model. As indicated in Table 5, the first segment visits more often because of their higher expected reading and writing utility conditioned upon visitation. This implies that reading and posting behaviors are correlated across users. We specify common scale parameters across segments because it is not clear why these scale factors should differ; (we find no significant difference when we do estimate them separately). Of note, conditioned on reading and posting utility, no parameters differ significantly across segments in the site visitation model. This suggests that heterogeneity in site visitation decisions are largely predicated on reading and posting utility.

7 Managing Site Sponsored Content: Quality and Quantity

In this section, we consider how the site can manage its own content development strategy. Importantly, we consider site sponsored content (SSC) that is *identical in construction* to

UGC except in quality. That is, we consider site sponsored forum posts that look identical to UGC from the perspective of the reader (e.g., the sponsorship is not revealed), and are indistinguisable except, potentially, in their quality. For an example of a sports forum, the site can invite expert sportsmen to post content. We focus on this type of content to ensure that the parameters in our UGC reading model are valid for inferring how SSC affects user behavior.

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 also consider two other cases pertaining to the *initial* level of user engagement at the start of the network, specifically to the role of content in getting a site to tip. With no content, there will be no readers attracted to the platform. With no readers, there will be no incentive to post. Hence, the site needs enough content to tip the network to a self-sustaining state. One way to do this 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 create a sufficient level sponsored content to tip the network – either with a large initial amount, a steady 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. If so, it may be better for the site to sponsor posts early and often than at a constant rate, because once the network tips, the site will no longer need to bear the expense of sponsoring.

While a definitive answer to which of these three strategies (initial user posts, initial sponsored posts only and regular sponsored posts) depends upon actual costs, one must first understand how these strategies affect network growth and exactly how much content is necessary to tip the network. Quantifying this amount is our aim in this section.

As sponsored and user content can differ in relative quality, this factor can also play a role in how each type of content affects site traffic. Hence, we first overview our approach to accommodate differences in site and user quality, and then proceed to our counterfactual scenarios.

7.1 The Effect of Site Sponsored Content on Reading

When a site sponsors additional content over and above that generated by users, readers have more content available. The reader will now choose the best posts across the entire set of UGC and SSC. Below, we derive the optimal total number of posts read and how they are allocated across the user and sponsored posts.

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

$$\begin{aligned} u(r^U_{it}, r^S_{it}) &= E\left(\sum_{k_s=K^S_t-r^S_{it}+1}^{K^S_t} Q_{[k_s]} + \sum_{k_u=K^U_t-r^U_{it}+1}^{K^U_t} Q_{[k_u]}|K^S_t, \Theta^S, K^U_t, \Theta^U\right) \\ &= \left(\left(U^S - L^S\right)\left\{\frac{K^S_t + 1/2}{K^S_t + 1} r^S_{it} - \frac{1}{K^S_t + 1}\frac{r^{S^2}_{it}}{2}\right\} + L^S r^S_{it}\right) \\ &\quad + \left(\left(U^U - L^U\right)\left\{\frac{K^U_t + 1/2}{K^U_t + 1} r^U_{it} - \frac{1}{K^U_t + 1}\frac{r^{U^2}_{it}}{2}\right\} + L^S r^S_{it}\right). \end{aligned} \tag{33}$$

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

$$u(r^U_{it}, r^S_{it}) = \alpha^S_1 r^S_{it} - \frac{\alpha^S_2}{2K^S_t} r^{S^2}_{it} + \alpha^U_1 r^U_{it} - \frac{\alpha^U_2}{2K^U_t} r^{U^2}_{it}, \tag{34}$$

where $\alpha^S_1 = U^S$, $\alpha^S_2 = U^S - L^S$, $\alpha^U_1 = U^U$ and $\alpha^U_2 = U^U - L^U$. Under the assumption of quadratic reading costs, the total cost of reading is given by $(\zeta_i + \kappa_{1it})\left(r^S_{it} + r^U_{it}\right) + 1/2\kappa_{2i}\left(r^S_{it} + r^U_{it}\right)^2$. Adding the user heterogeneity in the reading utility as in Equation (5), the optimal levels of reading user and sponsored content are then the solutions of the FOC where the marginal utility of reading is equal to the marginal cost:

$$\begin{aligned} \alpha^S_1 - \frac{\alpha^S_2}{K^S_t} r^S_{it} &= \zeta_i + \kappa_{1it} + \kappa_{2i}\left(r^S_{it} + r^U_{it}\right) \\ \alpha^U_1 - \frac{\alpha^U_2}{K^U_t} r^U_{it} &= \zeta_i + \kappa_{1it} + \kappa_{2i}\left(r^S_{it} + r^U_{it}\right), \end{aligned} \tag{35}$$

which also implies the marginal utilities from reading user and site sponsored content are equal. Notice that the marginal utility is the location of an additional post on the line of quality distribution and a reader will read all the posts whose quality is great than this post. Therefore, solving $r_{it}^S + r_{it}^U$ by Equation (35) is equivalent to selecting the number of highest quality posts from K_t^S and K_t^U combined.

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

$$\begin{bmatrix} r_{it}^{*S} \\ r_{it}^{*U} \end{bmatrix} = \begin{bmatrix} \left(\frac{\alpha_2^S}{K_t^S} + \kappa_{2i}\right) & \kappa_{2i} \\ \kappa_{2i} & \left(\frac{\alpha_2^U}{K_t^U} + \kappa_{2i}\right) \end{bmatrix}^{-1} \begin{bmatrix} \left(\alpha_1^S - \zeta_i - \kappa_{1it}\right) \\ \left(\alpha_1^U - \zeta_i - \kappa_{1it}\right) \end{bmatrix} \tag{36}$$

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

The ex ante theoretical effect of additional sponsored content on posting is ambiguous. On one hand, there is a competitive effect that lowers the likelihood that users' posts are read, thereby reducing users' incentives to participate in the site. On the other hand, increased content can generate more readership, thereby increasing the utility of posting and the resultant posts. In the next section, we consider this trade off explicitly and make recommendations regarding the site's participation levels.

7.2 Counterfactual Analyses

We begin with an analysis of the role of SSC in the state of network usage observed in our data, and then detail our analysis of the role of SSC and UGC in jump starting the network. We find the effect of site content on the currently observed state of the network is modest. However, we do find the site can play a substantial role in tipping the network at a low level of user engagement via the use of SSC.

7.2.1 SSC Conditioned on Current States

We simulate the effect of quality and quantity of SSC on UGC. Without loss of generality, we manipulate the quality of SSC by altering α_2^S, which we set to be smaller than α_2^U of UGC discussed in Section 7.1. The upper bound of SSC quality is represented by α_1^S, which we set to equal α_1^U. One reason we choose to manipulate α_2^S is that we fix $\alpha_1^U = 10$ and estimate α_2^U in estimation for identification. Notice that smaller α_2^S means SSC has higher mean quality ($\alpha_1^S - \alpha_2^S/2$) and less quality variance ($\alpha_2^{S^2}/12$) than UGC at the same time. We manipulate α_2^S to be either i) equal to the quality of UGC ($\alpha_2^S/\alpha_2^U = 100\%$) or ii) have a moderately higher quality than UGC ($\alpha_2^S/\alpha_2^U = 50\%$) and a substantially higher quality ($\alpha_2^S/\alpha_2^U = 10\%$). The resulting percentage change in average user-generated postings and number of visitors in equilibrium over a 70-day simulated sequence versus the levels of SSC is plotted in Figure 5.

Figure 5 demonstrates that SSC increases the number of visitors for a quality range of $\alpha_2^S/\alpha_2^U = 100\%$ and 50%. When the amount of sponsored content is at 10% level of the currently observed user-generated content, the increment in the visitation rate is about 5% for $\alpha_2^S/\alpha_2^U = 100\%$ and 10% for $\alpha_2^S/\alpha_2^U = 50\%$. Hence, the sponsored content arc elasticity is about .5 and 1 for these respective cases. The increment in UGC is much less pronounced thanks to the competition effect between UGC and SSC, gaining only about 1% and 2%, implying arc elasticities very close to zero.[22]

If the site can generate sufficiently large quantity of very high quality SSC ($\alpha_2^S/\alpha_2^U = 10\%$), the site can increase the visitation rate substantially. When $\alpha_2^S/\alpha_2^U = 10\%$ (i.e., all the SSC is within the top 10% of UGC quality), the site can increase the visitation rate by about 25%, for an elasticity of 2.5. However, the amount of UGC will increase only about 6% because of the competing effect of site posts on the likelihood that readers consume user posts. Between 7% and 10%, incremental SSC has a dampening effect on UGC. User posts are strongly affected by the shift of readers to higher quality sponsored posts. Overall, we conclude that there is potential to grow the network primarily with very high quality SSC.

[22]The seemingly oscillating behavior of the curves is due to sampling errors, as the mean number of postings is only the average of a 70-day simulated sequence.

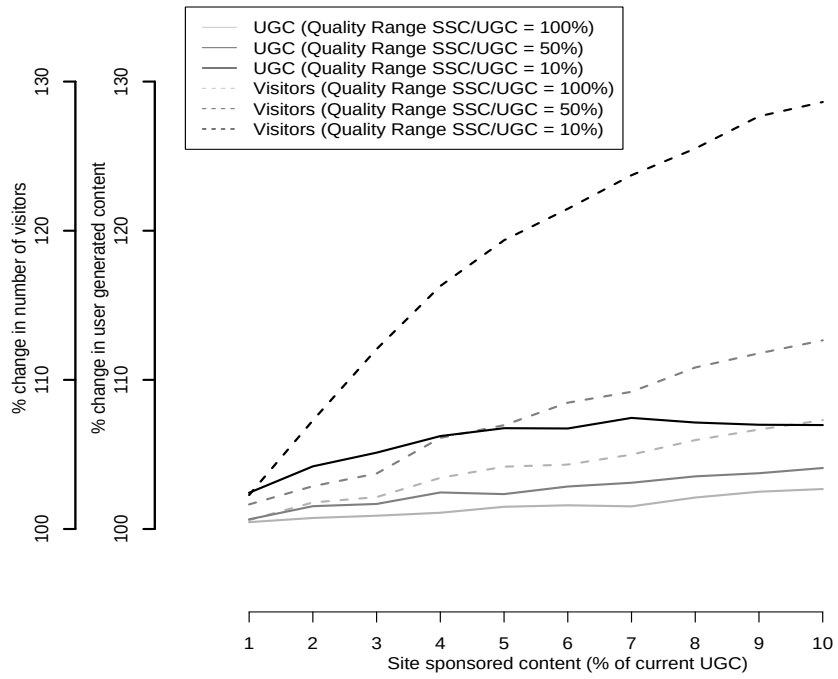

Figure 5: Effect of site sponsored content strategies on user-generated content and number of site visitor. The horizontal axis represents the percentage increase over the current average levels of UGC observed in our data. The vertical axis depicts the percentage changes in steady state site usage.

7.2.2 UGC Conditioned on Initial States

In the preceding section, we considered a counterfactual predicated on the current equilibrium observed in the data, one wherein the network has already become self-sustaining in terms of user participation because of the high levels of readers and posters. Another potential equilibrium for the UGC platform is in its infancy state, where individual postings, amount of reading and site visiting are all close to zero. This outcome self-sustains because low user-generated posting stock, set as the initial condition for the forum, attracts very few readers, which will further attenuate posting activity. Site activity converges to a degenerate equilibrium wherein user participation becomes negligible; hence the network implodes.

Therefore, an important policy simulation is to ascertain the tipping point (a.k.a critical

mass) of initial user-generated content, wherein any initial UGC lower than this critical point results in the collapse of the forum into an equilibrium of very low activity. We first consider the role of user stock on network tipping. Figure 6 demonstrates the relationship between the initial UGC stock, 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 the current number of visitors and amount of UGC we observe in the data). The critical user stock needed to tip the network is 10.7% of the current level of UGC. That is, when the initial UGC is below the 10.7% of the UGC we observe in the data, the forum will collapse into an equilibrium wherein site visitation rates reduce to only about 3% of the current level and the UGC falls to only about 0.5%. In contrast, when the initial UGC stock is about 10.7% or greater, the site will reach 100% of the current activity.

It is worth noting that modeling rational expectations profoundly affects the tipping point of the network. When beliefs about the participation of others is not allowed to evolve with changes in the system, we find 19.0% of the current observed levels of UGC is needed to tip the network.

7.2.3 SSC Conditioned on Initial States

Next, we consider the role of sponsored content 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 stock of posts. This strategy is analogous to sponsoring posts at the start of the network and then stopping this practice. 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.

Results suggest that SSC equal to 9.3% of the currently observed amount of UGC is sufficient to tip the network when the quality range ratio $\alpha_2^S/\alpha_2^U = 100\%$ and that 1% is sufficient when the quality range ratio is much higher, $\alpha_2^S/\alpha_2^U = 10\%$.

Next, we contrast these results to a case wherein, instead of initial posts only, the site continues sponsoring posts on a regular basis, and thus changes users' rational expectations regarding future SSC. The results of this analysis are reported in Figure 8.

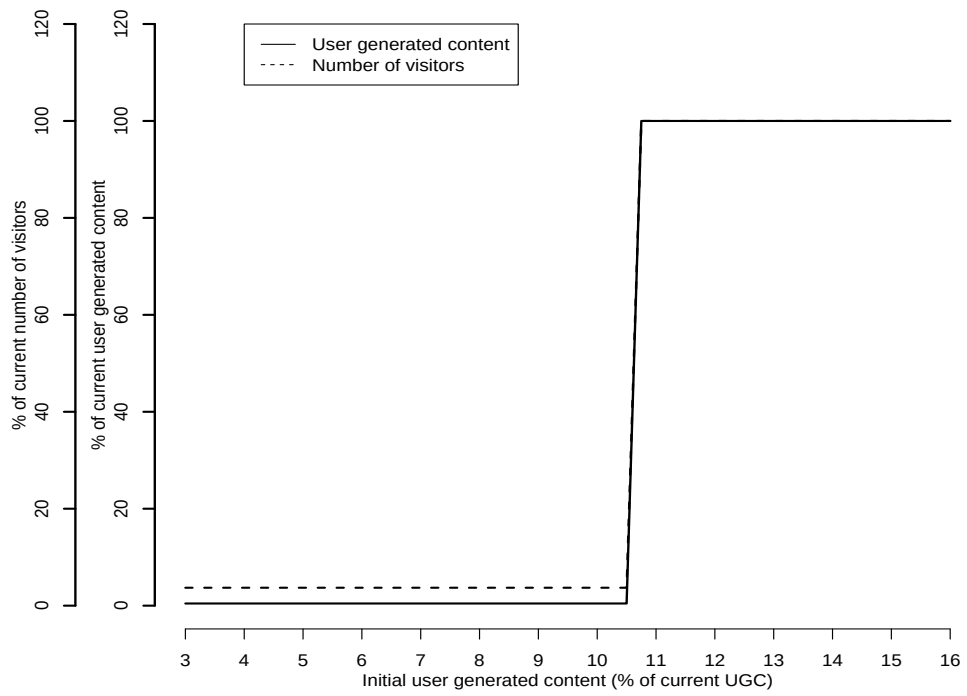

Figure 6: The critical point of initial UGC. The horizontal axis represents the percentage of initial user stock as a fraction of the current average levels of UGC observed in our data. The vertical axis depicts the steady state site usage.

As evidenced in Figure 8, the effect of a steady increase in sponsored content tips the network at a level equivalent to 8% of the currently observed UGC when site and user contents are equal in quality. The tipping point decreases further to 1% when the quality is very high ($\alpha_2^S/\alpha_2^U = 10\%$). The low tipping rates arise because users perceive a steady stream of high quality content availability. Hence, we find it is possible for a site to tip the network with a relatively small number of high quality posts, so long as the site is committed to sponsoring posts for an indefinite period of time.

Finally, we combined the two site strategies; initial SSC stock and steady SSC stock to assess potential synergies in inducing the network to tip. Of interest is whether these approaches are complements or substitutes. If the former, it suggests that both should be used. If the latter, it suggests one or the other should be used. Initial and continuing SSC

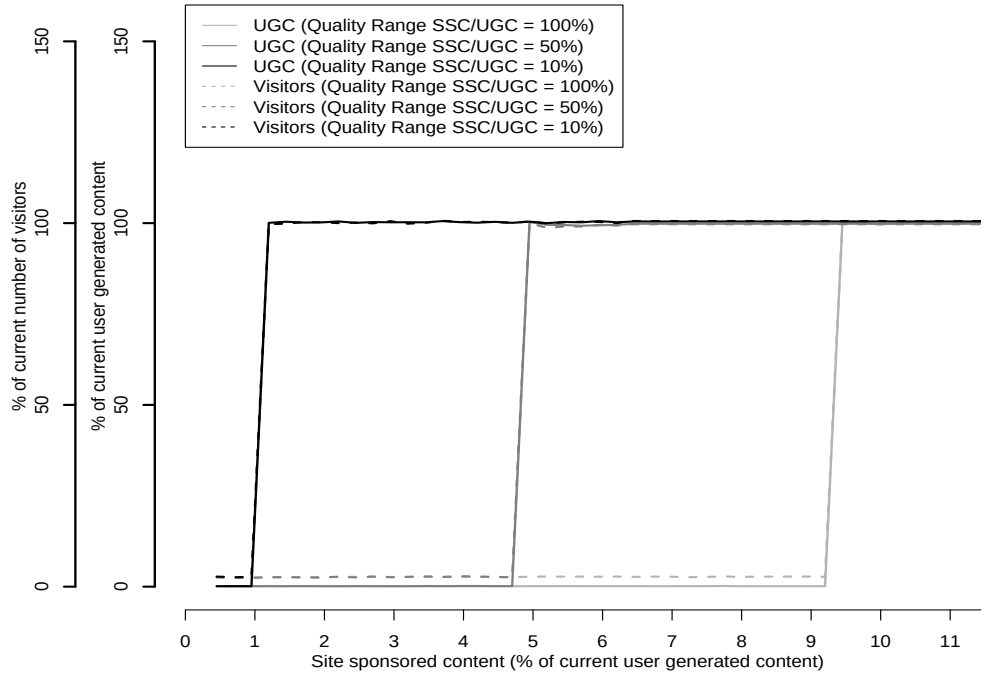

Figure 7: Effect of initial sponsored content strategies when initial UGC is set to zero. The horizontal axis represents the percentage of initial SSC as a fraction of the current average levels of UGC observed in our data. The vertical axis depicts the steady state site usage.

are substitutes; using both is redundant. Figure 9 presents the results.

Results suggest that the two strategies are substitutes, and that there is no synergy. Thus, it appears that the two approaches are somewhat redundant and that the firm should pick one strategy or the other to tip the network, depending on the relative long-term costs of each.

7.2.4 Site Content Management Strategies

Collectively, our counterfactual analyses indicate that site strategies have limited impact on the current steady state level of user engagement, but a more profound impact on ensuring the network takes off.

In terms of jump starting the network, we contrast three strategies: i) enlist users to post, perhaps through marketing via targeted advertising or incentives, ii) the site sponsoring

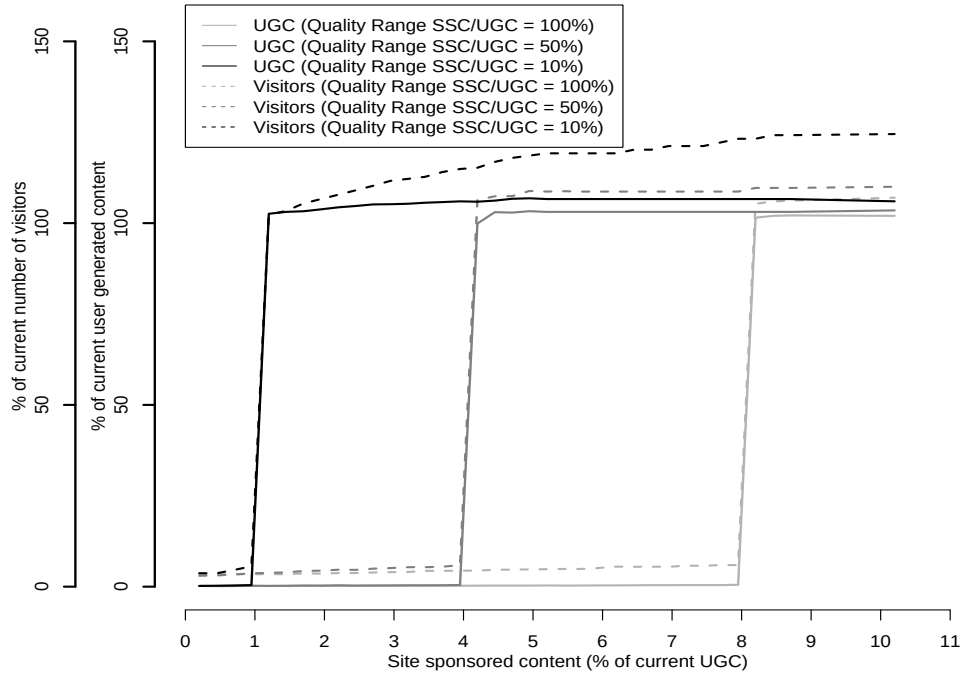

Figure 8: Effect of regular sponsored content strategies when initial UGC is set to zero. The horizontal axis represents the amount of SSC stock per period as a fraction of the current average levels of UGC observed in our data. The vertical axis depicts the steady state site usage.

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 post decreases. When the site sponsors posts instead, the aggregated value of these posts remains high enough to attract readers, but the individual user level stock is sufficiently low that their incentive to post is greater (their marginal benefit is higher). Because of this, the site tips more quickly.

The preferred strategy would be incumbent upon the relative cost of each. Contrasting whether the site should jump start with just initial posts (Figure 7) or a constant stream of posts (Figure 8), it is clear the latter is more effective at tipping. However, the cost of the

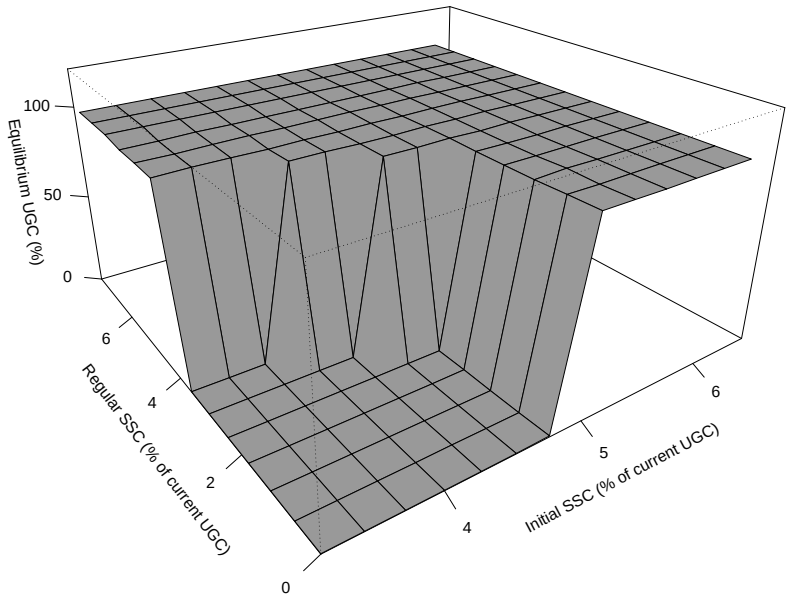

Figure 9: Interaction between initial and regular SSC for tipping when UGC is set to zero. The right horizontal axis represents the initial SSC as a percentage of the current average levels of UGC observed in our data. The left horizontal axis represents the regular SSC per period as a percentage of the current average levels of UGC. The vertical axis depicts the steady state site visitation rate as a percent of the maximum steady state level.

latter strategy is borne each period rather than once. This fact would imply a tendency to favor a strategy of sponsored posts early for firms with low discount rates. When contrasting whether to use initial user or initial sponsored post stock, much depends on the relative cost of attracting each. For example, were the site able to sponsor high quality posts at a reasonable cost in a short period of time, this would favor the strategy of such posts at the network's start, and then withdrawing completely after the UGC tips to the favorable equilibrium.

8 Conclusions

Recent advances in technology and media have enabled user generated content sites to become an increasingly prevalent source of information for consumers as well as an increasingly relevant channel for advertisers to reach users of these sites. Hence, the factors driving the

use of these networks is of a topical concern to marketers. In this paper, therefore, we consider how content, readership and site policy drive the evolution of content and readership on these sites.

Given our goal is to develop prescriptive and theoretical insights regarding user engagement on user generated content platforms, we build upon the existing literature on social participation by developing a dynamic structural model to explore these effects. Individual reading behavior is developed from a model of information search that relates reading to the overall level of content on the site. Individual content generation is assumed to reflect the utility that participants receive from the number of others reading the posts. Underpinning these two behaviors are users' beliefs regarding how the aggregate amount of content and readership on the platform evolve. These beliefs stem from the rational expectations equilibrium model whereby the evolution of aggregate reading and content states are 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 user generated content. 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 of Krusell and Smith (1998) along multiple dimensions, including i) enabling a single unit of supply (posts) to be consumed *concurrently* by many (readers), ii) enabling agents to *both* produce and consume, iii) accommodating both *continuous* and discrete behaviors, iii) applying computational advances to enhance the *scale*, including embedding approximate aggregation into an MPEC estimator. As a result, our approach facilitates the computation of a rational expectations equilibrium in the face of a large number of heterogeneous agents. As such, our advances could prove useful in other contexts in marketing and economics wherein firms face heterogeneous consumers. For example, heterogeneous learning about new consumer products can affect how prices evolve, and consumer may anticipate and react to such changes (Narayanan and Manchanda 2009). Initial estimates of our model of UGC demonstrate that the indirect network effect or aggregate reading on posting and aggregate posting on reading are both significant.

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

On a substantive domain, we consider a number of policy prescriptions to advise the site. First, we consider the role of their sponsored content on user participation in a mature network. On the one hand, sponsored posts attract more readers, thereby growing the network. On the other hand, these posts are competitive with other users' posts. Overall, we conclude that the former effect predominates and the site can increase visitation by 10% by increasing content by 10% if sponsored posts are of sufficiently higher quality. Next, 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 (wherein there is sufficient content to attract readers and sufficient readers to attract content) with either 10.7% of the current content if users are incentivized to post or 9.3% if the site posts. The difference can be ascribed to users diminishing marginal returns to posting – when the site jump starts the network, users have lower posting stock and thus a higher incremental value of posting. An alternative site strategy is to increase its steady state stock as opposed to its initial stock. This difference is tantamount to contrasting a strategy of continual posting to one wherein the site exits once the network has tipped. It takes only 8% of content to induce tipping when users believe the site will post on a continual basis. Which strategy is most effective (initial user posts, initial sponsored posts, or regular sponsored posts) is incumbent upon the relative costs of each strategy. However, we note that a strategy of jump starting a network means a site can stop bearing the cost of sponsoring posts once the network tips. Unless the incremental costs of posting are highly convex, the site can accordingly benefit from a jump start strategy. Indeed, when sponsored post quality is sufficiently high, it only takes about 1% of the observed posting levels in the mature network to tip it. Perhaps this is one key reason that this latter approach was the one chosen by `Soulrider.com` to grow its network.

Our study offers evidence of the efficacy of this strategy.

Several opportunities for extensions are present. First, the potential for competition exists for forum sites and extending our work to a duopoly context would be of interest (Zhang and Sarvary 2011). Second, it would useful to extend our model to capture heterogeneity in content information in order to explore what information is most relevant in increasing site engagement. Related, the potential exists that certain lead content creators generate large followings and measuring the effect of lead users is of practical interest. While sites generally consider such participation to be positive, it is also possible for the content to compete with others and actually reduce site participation. Of note, the last two extensions potentially involve a considerable expansion of the state space and would also afford the opportunity to offer advances in numerical computation in order to become feasible. Finally, our analysis considers a site where the number of times a post is read is not shown to users and posts are not rated. While accommodating observed levels of reading is relatively straightforward (one can use the actual number of reads instead of the rational expectations for reads), the ratings of posts provide another incentive to post and would likely enter a joint posting-rating utility function. Moreover, the incentive to rate would need to be considered. Owing to the prevalence of sites with rated content, this is an interesting future direction.

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

References

Albuquerque, Paulo, Polykarpos Pavlidis, Udi Chatow, Kay-Yut Chen, Zainab Jamal, Kok-Wei Koh, Andrew Fitzhugh. 2010. Evaluating Promotional Activities in an Online Two-Sided Market of User-Generated Content. *SSRN eLibrary* .

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

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

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

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

Choi, Ki-Hong, Choon-Geol Moon. 1997. Generalized extreme value model and additively separable generator function. *Journal of Econometrics* **76**(1-2) 129 – 140.

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

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

Dichter, E. 1966. How word-of-mouth advertising works. *Harvard Business Review* **44**(6) 147–160.

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

Dubé, Jean-Pierre, Günter J. Hitsch, Puneet Manchanda. 2005. An Empirical Model of Advertising Dynamics. *Quantitative Marketing and Economics* **3** 107–144, 10.1007/s11129-005-0334-2.

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

Dubé, Jean-Pierre H., Jeremy T. Fox, Che-Lin Su. 2009. Improving the Numerical Performance of Blp Static and Dynamic Discrete Choice Random Coefficients Demand Estimation. *SSRN eLibrary* .

Geweke, John, Micheal P. Keane. 1997. Mixture of normals probit models. *Research Department Staff Report 237, Federal Reserve Bank of Minneapolis* .

Ghose, Anindya, Sang Pil Han. 2011. A Dynamic Structural Model of User Learning on the Mobile Internet. *SSRN eLibrary* .

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

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

Huang, Yan, Param V. Singh, Anindya Ghose. 2011. A Structural Model of Employee Behavioral Dynamics in Enterprise Social Media. *SSRN eLibrary* .

Iyengar, Raghuram, Christophe Van den Bulte, Thomas W. Valente. 2010. Opinion leadership and social contagion in new product diffusion. *Marketing Science* mksc.1100.0566.

Kamakura, Wagner A., Gary J. Russell. 1989. A probabilistic choice model for market segmentation and elasticity structure. *Journal of Marketing Research* **26**(4) pp. 379–390.

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

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

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

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

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

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

Miranda, Mario S., Walter D. Fackler. 2002. *Applied Computational Economics and Finance*. The MIT Press, Cambridge, MA.

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

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

Narayanan, Sridhar, Puneet Manchanda. 2009. Heterogeneous learning and the targeting of marketing communication for new products. *Marketing Science* **28** 424–441.

Nardi, Bonnie A., Diane J. Schiano, Michelle Gumbrecht, Luke Swartz. 2004. Why we blog. *Commun. ACM* **47** 41–46.

Nov, Oded. 2007. What motivates wikipedians? *Communications ACM* **50** 60–64.

Ransbotham, Sam, Gerald C. Kane, Nicholas H. Lurie. 2012. Network characteristics and the value of collaborative user-generated content. *Marketing Science* .

Rust, John. 1987. Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher. *Econometrica: Journal of the Econometric Society* (5) 999–1033.

Rust, John. 1994. *Structural Estimation of Markov Decision Processes*. Amsterdam: Elsevier Science.

Shriver, Scott K., Harikesh S. Nair, Reto Hofstetter. 2013. Social ties and user generated content: Evidence from an online social network. *forthcoming, Management Science* 1–.

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

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

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

Sundaram, D. S., Kaushik Mitra, Cynthia Webster. 1998. Word-of-mouth communications: A motivational analysis. Joseph W. Alba, J. Wesley Hutchinson, eds., *Advances in Consumer Research*, vol. 25. Association for Consumer Research, 527–531.

Yao, Song, Carl F. Mela. 2008. Online Auction Demand. *Marketing Science* **27**(5) 861–885.

Zhang, Kaifu, Theodoros Evgeniou, V. Padmanabhan, Emile Richard. 2011. Content contributor management and network effects in a ugc environment. *Marketing Science* **Forthcoming** 1–.

Zhang, Kaifu, Miklos Sarvary. 2011. Social media competition: Differentiation with user generated content. *Working Paper, INSEAD* 1–.

Web Appendix

A Aggregate Reading

Here we show the expected amount of reading per posting y_t define in equation (9) can be closely approximated by the observed amount of reading per posting. The expected amount of reading of a given user, we obtain the aggregate expected amount of reading by all users,

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

where M is the total number of users.[23] When we apply the latent segment model, the expected amount of reading of any user i is

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

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

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

The observed total amount of reading which is denoted by $\tilde{R}_t$ is defined by

$$\tilde{R}_t = \textstyle\sum_{i=1}^{M} n_{it} r_{it} = \displaystyle\sum_{i=1}^{M} \sum_{j=1}^{J} n_{it} I\left(\zeta_i = \bar{\zeta}_j\right) \frac{\alpha_1 + \bar{\zeta}_j - \bar{\kappa}_{1j} w_t}{\alpha_2 / K_t + \bar{\kappa}_{2j}} \nu_{it},$$

so it is obvious that

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

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

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

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

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

B Rational Expectations

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

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

2. Guess the values for ω_0^K, ω_1^K, ω_2^K and $\omega_0^y, \omega_1^y, \omega_2^y$ in equations (27) in Section 3.6 and (28).

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

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

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

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

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

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

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

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

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

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

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

C Model Estimation

C.1 Estimating the Reading Model

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

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

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

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

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

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

C.2 Estimating the Posting Model

One key component of estimation is to approximate the expected value functions in equation (17). This task is nontrivial for our model, because our state variables are mostly continuous with a wide support. Moreover, the control variable can take high-order discrete values. For

this reason, we use Chebyshev approximation to approximate the expected value functions as described in (Dubé et al. 2009; Miranda and Fackler 2002). Chebyshev approximation uses polynomial interpolation to approximate the expected value functions:

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

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

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

To compute the right-hand side of the above equation, we need to numerically evaluate an indefinite integral with respect to state transition probabilities of aggregate stock of posting. Since we use a normal distribution to model the probabilities, the Gauss-Hermite quadrature can be used to approximate the integration in the Bellman equation above Miranda and Fackler 2002. The Gauss-Hermite quadrature allows us to evaluate the integrand at fewer points than, for example, a Monte Carlo integration.

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

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

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

The joint likelihood of reading and posting for all individuals is then[24]

$$\left\{\prod_{i=1}^{M}\sum_{j=1}^{J} p_j \prod_{t=1}^{T} \text{Exponential}\left(r_{it}|\frac{\alpha_1 - \bar{\kappa}_{1j}w_t - \bar{\zeta}_j}{\alpha_2/K_t + \bar{\kappa}_{2j}}\right) \frac{\exp\left(u(a_{it}|s_{it}) - \bar{c}(a_{it}|s_{it}) + \tilde{EV}_j(s_{it},a_{it})\right)}{\sum_{a'_{it}\in A}\exp\left(u(a'_{it}|s_{it}) - \bar{c}(a'_{it}|s_{it}) + \tilde{EV}_j(s_{it},a'_{it})\right)}\right\}. \tag{A4}$$

The direct MLE approach (e.g., Kamakura and Russell, 1989) is applied to estimate the parameters. To compute the standard errors of parameter estimates in the posting model,

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

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

C.3 Estimating the Site Visitation Model

Lastly, we have the likelihood function for site-visitation data following equation (26):

$$\left\{ \prod_{i=1}^{M} \sum_{j=1}^{J} p_j \prod_{t=1}^{T} \left[P(n_{it} = 1|s_{it})^{n_{it}} P(n_{it} = 0|s_{it})^{1-n_{it}} \right] \right\}, \tag{A5}$$

which is also estimated by MLE.

D Theoretical Implications by Simulation

In this section, we explore some of the theoretical properties of our model. Specifically, we assess i) convergence to the defined rational expectations equilibrium in Section 3.6 and ii) how the model's parameters and exogenous states influence the network's user content and readings in equilibrium. This analysis considers the role of initial content on network size, the effects of reading and posting costs and content stock decay on content generation, and the self-fulfilling prophecies under rational expectations.

D.1 Simulation Design

We simulate 2 segments of 3000 and 6000. We let both segments have the same cost of reading and heterogeneous costs ($\bar{\xi}_j$) of content generation. To simply the simulation, we assume there is no cyclical effect ($\bar{\tau}_j = 0$ and $\bar{\kappa}_{1j} = 0$). The reading cost parameters $\alpha_1 - \kappa_1 = 0.1$, $\alpha_2 = 1$ and $\kappa_{2j} = 0.0015$ imply a posting stock of $K_t^U = 10,000$ will induce an individual user to read 62.5 different postings per period. We let the cost of posting for Segment 1 be $\bar{\xi}_1 = 0.1$ and segment 2 be $\bar{\xi}_2 = 5$. Note that $\bar{\xi}_2$ is 50 times of $\bar{\xi}_1$, which implies Segment 2 has a much higher cost of posting and hence users in Segment 2 are likely to post much less than those in Segment 1. Indeed, we find in equilibrium a user in Segment 2 writes only

about 2 postings in 100 periods whereas a user in Segment 1 writes about 350 posting in the same periods on average. We set the posting utility parameter $\gamma = 0.5$.

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

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

D.2 Simulation Results

D.2.1 Initial Individual Stock

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

In Figure 10, we plot the equilibrium path of the aggregate user generated postings (UGC) after the rational expectations equilibrium is achieved. We can see that the first set of initial values (solid curve) and the second (dashed curve) converge to the same steady-state aggregate UGC with small, random variations. We also find the same equilibrium parameter values in the equations(27) and (28) in Section 3.6. The UGC reaches the steady state after only about 10 periods.

Based on the theoretical model in Section 3.3, we expect that not only the aggregate UGC converges (shown in Figure 10), but also the distribution of individual posting stocks would be constant in the steady state as well. Figure 11 shows the distribution (histogram)

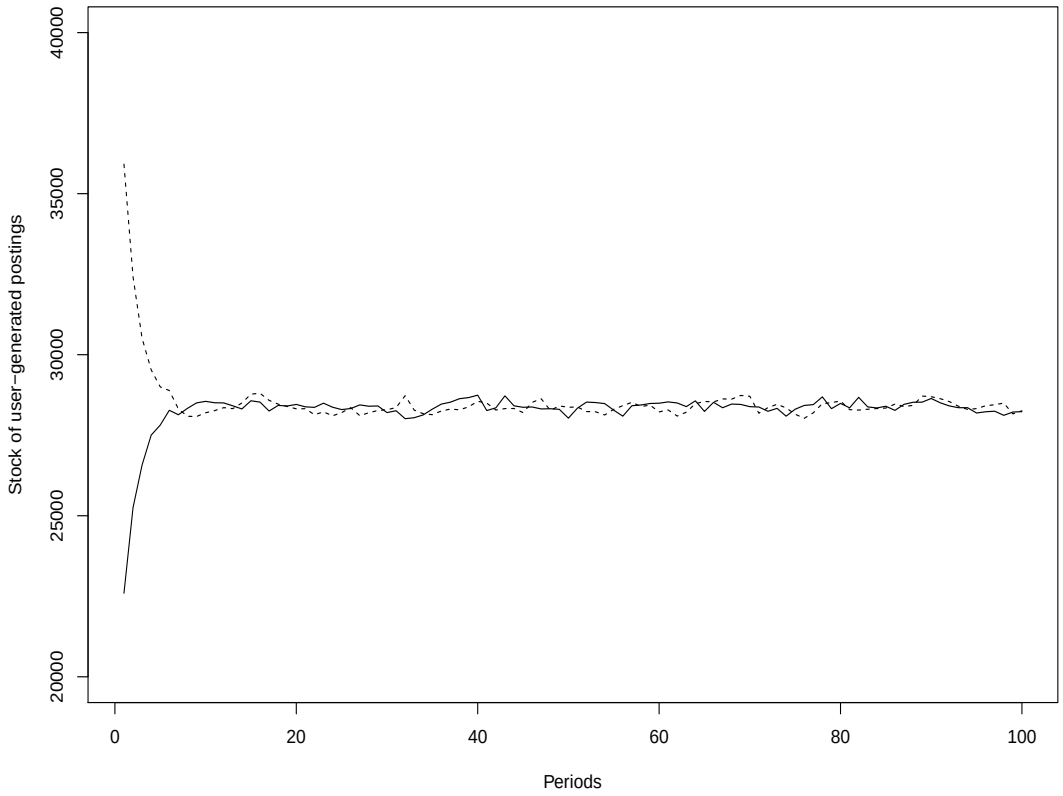

Figure 10: Convergence of aggregate user-generated posting stock (UGC) to the steady state from 2 different starting values.

of individual posting stocks of the two segments of site users in period 50 and 100, when the UGC has already reached the steady state. These histogram plots confirm our conjecture that these distributions are indeed invariant over time.

D.2.2 Degenerate Equilibrium

A potential equilibrium of our model is that all postings, reading and visitation are zero. That is, the network will never expand unless some shock or intervention enables the network to tip from a non-zero state. For example, an extremely low user- posting stock can cause the low reading and visiting rate, which can in turn cause even lower posting activity. In order to test this conjecture, we select a set of very low initial values for posting stocks: 0.1 for both Segments 1 and 2. The dashed curve in Figure 12 demonstrates the result of this

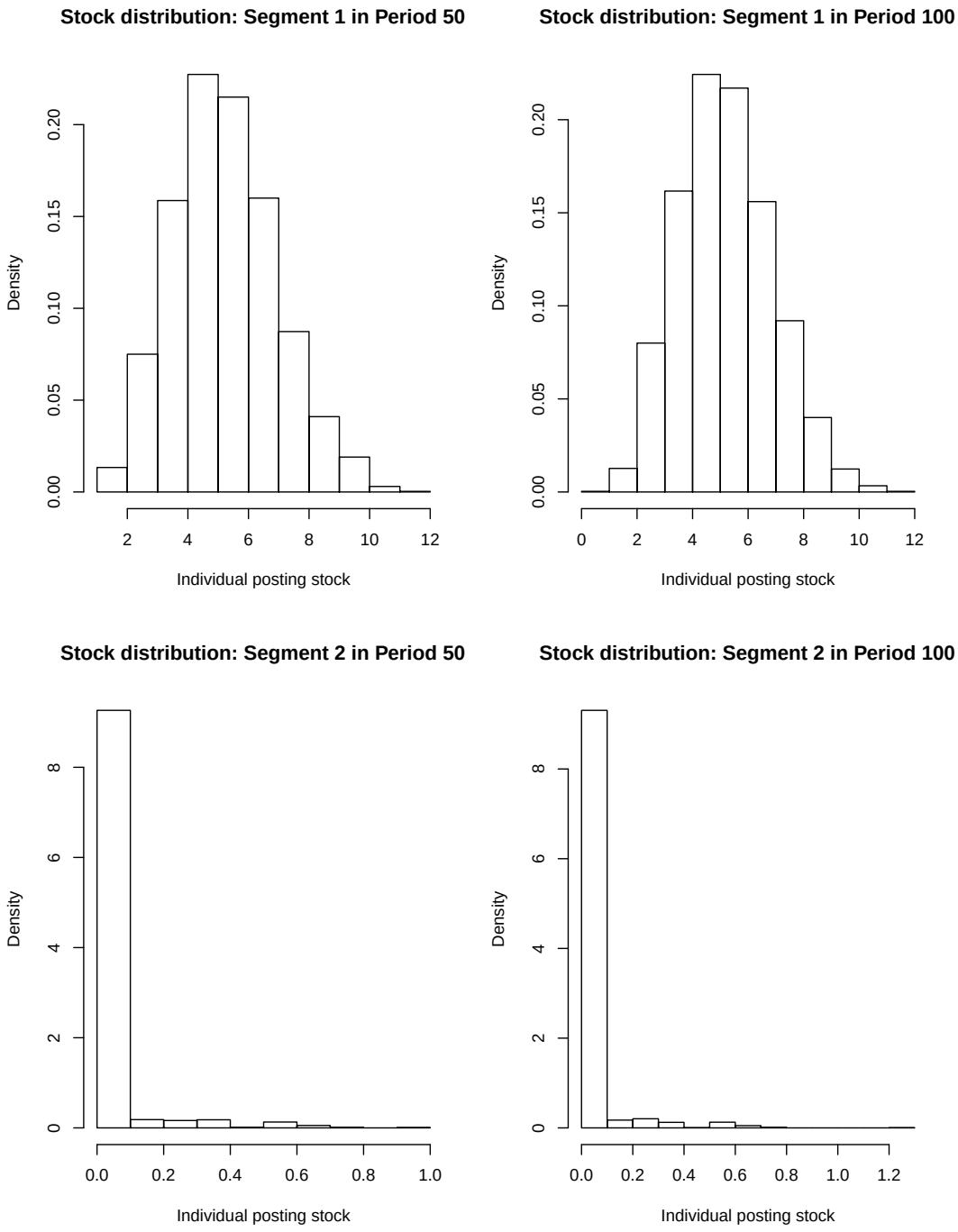

Figure 11: Distributions of individual user's posting stocks of the two segments defined in Section D in steady state.

simulation which converges to the trivial equilibrium.

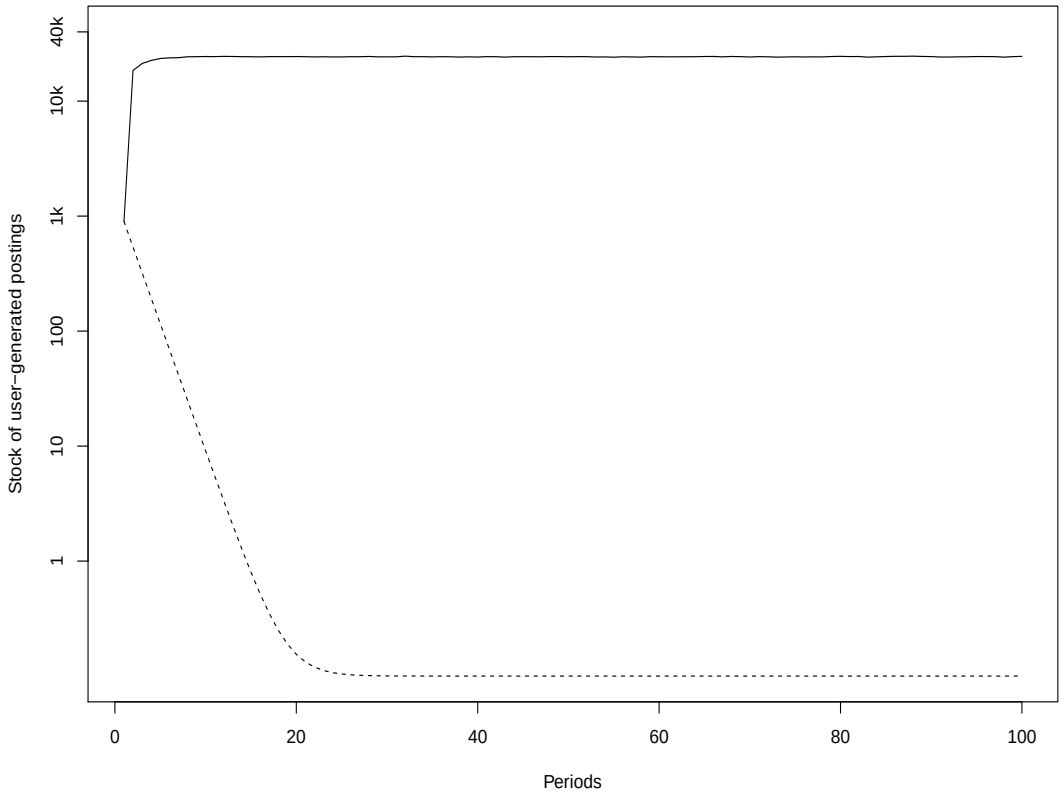

Figure 12: Convergence of the aggregate user-generated posting stock (UGC) to two different steady states from a common starting value when the initial individual posting stock is 0.1 and (i) the site sponsored content (K_t^S) has the means equal to $20,000$ (solid curve) and (ii) $2,000$ (dashed curve).

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

The decay parameter ρ of site's postings implies two opposite effects on user posting activity. First, a lower decay rate (higher ρ) means a post is more likely to be seen in the future, so a user has the incentive to post more. This also raises content available for readers thereby increasing site visitation. However, higher ρ makes posting more "durable" and hence increases the aggregate posting stock and decreases the rate of reading per posting, which could cause a user to post less. The net effect of ρ is not clear directly from the utility

function because a closed-form derivative of the utility with respect to the decay parameter cannot be easily derived. Therefore, we discretize the space of the decay parameter ($\rho \in [0, 1]$) to ten equally spaced grid points (0.1, 0.2, ..., 0.9) and simulate the content and reading given these values.

In Figure 13, we depict the relationship between the decay parameter and the average number of postings per period per user in Segment 1 (solid curve) based on the simulation results. Figure 13 also indicates the relationship between the decay parameter and the average reading per posting y_t (dashed curve). From Figure 13, a higher decay parameter will *ceteris paribus* cause lower average reading per posting thanks to the competitive effect of more durable postings. However, the average number of postings per user increases when the decay parameters ρ increases from 0.1 to 0.4 and decreases when ρ is above 0.5. This result is due to the two opposite effects of ρ on user activity.

D.2.4 Self-fulfilling Prophecies

Owing to the formation of expectations regarding the aggregate state transitions in Equations (28) and (29), content generation and reading decisions are incumbent upon future beliefs. Of interest is the possibility that these beliefs become self-reinforcing. This issue can be explored by shocking these beliefs in the short-term (by varying the initial states and variances in the state transition equations) and the long-term (by varying the regression coefficients in the state transition equations) seeing how the evolution of content generation changes.

In order to test whether shocking short-term beliefs can lead to different long-term behaviors, i.e., converging to different equilibria of the model, we reset the initial belief about the aggregate UGC stock to 5%, 25%, 50%, 150%, and 200% of the actual stock and simulate the rational expectations equilibrium following the algorithm in Section 3.6. We find all these simulations converge to the same equilibrium which has the same levels of mean UGC and number of visitors as in the observed data. We also reset the initial belief about the variance in the state transition equation for the aggregate UGC to 25%, 50%, 150%, 200% and 300% of the value estimated from the real data. All these simulations again converge to the original equilibrium. Hence, we conclude that shocking short term beliefs will not lead to self-fulfilling behavior.

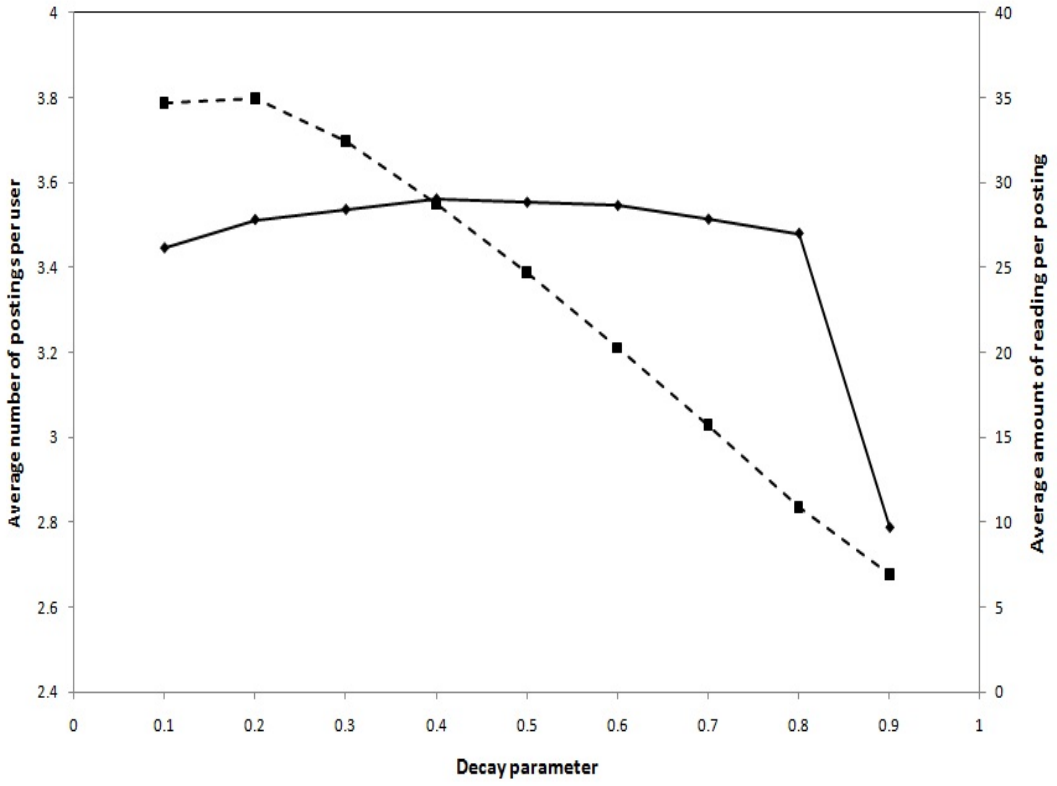

Figure 13: Average number of postings by individual users in steady state vs. the decay parameter ρ (solid curve) and the average reading per posting y_t vs. the decay parameter ρ (dashed curve).

To evaluate whether erroneous long-term beliefs about the transition rule of aggregate UGC can lead to different equilibrium, we set the initial value of the coefficient ω_1^K in equation 29 to 0.1, 0.2,...,0.9 and simulate their corresponding equilibria. We find they converge to the same equilibrium in which the auto-regressive coefficient ω_1^K is 0.84. Therefore, erroneous long-term beliefs about the transition rule will not lead to self-fulfilling behavior. Note that the rational expectations equilibrium in our model is similar to that by Krusell and Smith (1998), who also found the absence of self-fulfilling behavior in their model.