Incentive and Equilibrium of User Content Generation — A Theoretical and Empirical Study of Twitter

Huaxia Rui Andrew B. Whinston

University of Texas at Austin

ruihuaxia@mail.utexas.edu

September 15, 2009

Abstract

With the advent of broadband Internet, user generated content has prospered during the last few years. Various online communities like Wikipedia, Twitter, Flickr, and Youtube are offering unprecedented opportunities for ordinary people to access a huge amount of content and to publish their own content for a global audience. Users in such communities interact with each other by viewing others' content (consumption) and generating their own content (production). We model the pattern of user content generation in such large online communities as the equilibrium of a game where users make consumption and production decisions to maximize their utility. Unlike most of the previous literature where contributing content is deemed as a purely cooperative behavior like the provision of public good, we explicitly model users' private incentives of contributing content by taking into account users' preference for attention in the utility function. We show that as the community size grows large enough, all Nash equilibria converges to one limit equilibrium where the segmentation of the community could be characterized very neatly. Furthermore, if users' utility functions satisfy a certain linearity condition, as the community size goes to infinity, with probability 1, users self-select into either the group of content producers or the group of content consumers. We also show that the macro-level content consumption and production of such partition equilibrium is stable which suggests that massive content generation in large online communities is sustainable. Based on the theoretical model, we propose three hypotheses which are then tested using data collected from Twitter. Overall, our model is well supported by the empirical study. These results give insights into the pattern of user content generation in large online communities. They also provide a new angle of the free-riding phenomenon in online communities.

1 Introduction

Internet has become an indispensable part of people's daily life. One of the most important advantages of the Internet is that it has enabled people to access a huge amount of content at a very low cost. Part of this is because of powerful search engines like Google, Yahoo! and Bing which are freely available to everyone. Equally important is the abundance of free content available on the Internet, an ever growing proportion of which is now generated by ordinary Internet users. For example, Wikipedia, the largest and most popular general reference resource on the Internet, has over 13 million articles as a result of contributions by volunteers. On Flickr, over 3.6 billion images have been published by users all over the world. According to Youtube Fact Sheet, people are watching hundreds of millions of videos a day on YouTube and uploading hundreds of thousands of videos daily. On Twitter, a microblogging site that allows its users to broadcast short messages to their followers, people can find opinions and information on a broad range of topics posted by their peers almost in real time.

Despite the difference in the content format (text, image, video, etc), they are all produced by large number of ordinary Internet users, rather than by only a few publishers and television networks as in traditional media. This ongoing shift to social media means that user generated content is now playing an unprecedented role in people's life and will probably completely alter the way content is generated in our society in the future.

The fascinating phenomenon of user content generation also poses interesting research questions. For example, although it is quite reasonable that users consume others' content because they get benefits from content consumption, it is not yet clear what motivates people to contribute content. As is often reported, only a small proportion of users actually contribute content while most of users are ”inactive” in the sense that they don’t contribute any content at all. For example, on Wikipedia, the top 15% of the most prolific editors account for 90% of Wikipedia’s edits. In a recent article on the blog of Harvard Business School, Bill Heil and Mikolaj Piskorski [1] found that the top 10% of prolific Twitter users accounted for over 90% of tweets and they suggest that Twitter resembles more of a one-way, one-to-many publishing service.

Obviously, there is a missing link from the question of why users contribute content to the question of how does users’ usage pattern look like in an online community. The aim of this paper is to provide such link by modeling users’ incentive to consume and produce content and study how the incentive structure shapes the outcome of the interplay among a large number of users.

Besides the above questions, we are also interested in whether online communities supported by user generated content are sustainable in the sense that they are robust to small perturbations of user behavior. We take the above as our research questions in this paper and try to answer them through both theoretical modeling using game theory and empirical testing with data collected from Twitter. Our basic assumption for the model is that the community size is very large so we mainly focus on the asymptotic properties of equilibria.

First, we model each user as a rational agent who maximizes his or her utility from consuming and producing content. The consumption of content gives value to a user because he/she gets information, insight, or pleasure from the content. The more content available, the more value he or she will get from consuming content. On the other hand, users obtain utility from content production because he or she gets attention, publicity, vanity or ego gratification from peer recognition [1]. The more users consuming his or her content, the more value he or she derives from the content production. We will discuss users' motivation to produce content in detail in Section 2 and 3. Users may have different weights for the two parts of utility and they may also be heterogeneous in terms of productivity. Given the limited endowment of time, each user makes simultaneous and deliberate decisions on how much time to spend on content consumption and how much to spend on content production. Their decisions are interwinded since a user's utility from consumption depends on others' production decisions while a user's utility from production depends on others' consumption decisions. We explore the equilibrium outcome of this game by characterizing how the community is segmented into content consumers, content producers, and content prosumers who are users engaging in both content production and consumption. Interestingly, all Nash equilibria of the game converges to the same limit equilibrium as the community size grows to infinity which greatly simplifies our analysis of the usage pattern on Twitter in the empirical part of the paper. We also identify the existence of a special equilibrium called partition equilibrium which occurs when users' utility function satisfies a certain linearity condition. In such an equilibrium, users self-select themselves into either the group of content consumers or the group of content producers but not both. These results suggest that there is a tendency

[1] For example, Twitter users who produce high quality content attract lots of followers and get a lot of publicity. See http://wefollow.com for instance

for specialization of users' utility maximization strategies. They also shed some new light on our understanding of the role of those so-called "inactive" users in online communities.

To understand the sustainability of user content generation, we extend the model to the dynamic setting where users choose endowment of time based on their opportunity costs. In order to keep the model tractable, we focus on the partition equilibrium. We find that the dynamic system that characterizes the content demand and supply at the macro-level has a non-trivial asymptotically stable equilibrium point which suggests that the online community is rather stable under the partition equilibrium since the macro-level content demand and supply is stable and each user only responds to the macro-level variables. This result provides theoretical support to the sustainability of user content generation in online communities.

Second, we collect data from twitter.com, a popular social networking site, to test three hypotheses derived from the theoretical modeling. Our first hypothesis says that users who value attention more tend to produce more content. We support this hypothesis with about 1 million Twitter user profiles. We notice that about 20% of the Twitter users put URL on their profiles to promote something [2] while the other 80% users do not. Rather than a random phenomenon, we believe this is the result of a self-selection process where users who highly value attention put URL on their profiles while users who do not care so much about attention do not put URL on their profiles. The second hypothesis says that more capable Twitter users (in terms of producing content) will produce more content. Since users' capability is unobserable, we use the number of followers a user has as a proxy variable to

[2] For example, their personal blog site, the organizations they belong to, etc.

his or her capability. Our third hypothesis is based on an econometric model characterizing the distribution of contribution frequencies which is consistent with the data.

The paper is organized as follows. The next section reviews the relevant literature. Section 3 presents our theoretical model where we set up the model and characterize the segmentation of the community. Section 4 presents our empirical test of the model. We conclude in Section 5 with a discussion of the findings, their implications and limitations, and suggestions for future work. All proofs are omitted due to page limit.

2 Relevant Literature

There is a diverse literature related to our research questions, among which we will mainly introduce three streams of literature, economics, management, and computer science.

From traditional economics' point of view, contributing content in an online community is like the private production of public goods. Producing content is a cooperative behavior while consuming content without contributing content is a non-cooperative behavior. The puzzle is why so many users actively spend their valuable time providing content instead of free-riding others' contribution which will eventually lead to the collapse of online communities. Unlike offline communities where formal enforcement mechanism like contracts and agreements can be made, in online communities, people are only loosely connected and formal means are just infeasible. Economists have also studied informal enforcement mechanisms including personal enforcement and community enforcement. Personal enforcement involves retaliation to the non-cooperative agent by the victim who plays a cooperative strategy. Although the

Folk Theorem in the repeated game literature shows that any feasible outcome satisfying minmax condition could be achieved if there is frequent and long-term relationship among players [2, 3], in most large online communities, however, users seldom have such kind of long-term relationship. On the other hand, the community enforcement mechanism offers another explanation to why people cooperate [4]. Roughly speaking, under no information processing, cooperation may be sustained in the community because people fear that if they stop cooperating, non-cooperative behavior will "spread" like a disease to others, and this contagious process will eventually bring down the community which they all benefit from. Although such cooperative equilibrium seems to fit the context of online communities, it is quite fragile in the sense that small noise may cause the complete breakdown of cooperation in the community. Nowadays, large online communities often have tens of millions of users. It will be extremely difficult if not impossible to sustain such cooperative equilibrium.

Rather than assuming that contribution is purely a cooperative behavior that only benefits others, some economists and sociologists take another approach by arguing that contributing itself benefits the contributor. In the literature of economics of gift and charity, researchers suggest that people have a taste for giving. Andreoni [5] argues that "egoists" and "impure altruists" not only care about supplying public good, but also experiencing "warm glow" from having "done their bit." . Roberts, Hann and Slaughter [6] categorizes three factors that could motivate users to contribute to open source software (OSS) development, namely intrinsic, extrinsic, and internalized extrinsic factors. Intrinsic factors refers to the satisfaction and enjoyment obtained from creating and contributing [6, 7] while extrinsic factors refers to the incentive provided by external environments including organization rewards [8], career opportunities [6, 7, 9], etc. Internalized extrinsic factors refers to extrinsic motivations that are self-regulated instead of directly imposed by external environments, like reputation [10] and status seeking [6]. It may seem mysterious that OSS developers donate their valuable time and effort to develop open source software, which could be viewed as a special type of user generated content. However, as is argued by Lerner, Pathak and Tirole [11], open-source software developers may get some short- or long-run benefits. For example, a programmer may find intrinsic pleasure, get ego gratification from peer recognition, attract potential future employers, etc. Whether it's the intrinsic factors, extrinsic factors, or internalized extrinsic factors that motivate a user to contribute, it is reasonable to assume that in most cases, more attention will lead to higher value to users who contribute. This is also consistent with Lerner, Pathak and Tirole's suggestion that the more visible the performance to the relevant audience(peers, labor market, and venture capital community), the stronger such benefits will be. In this paper, we do not go into the underlying psychological and sociological mechanisms of why people contribute content in online communities. Rather, we assume that these underlying mechanisms manifest themselves through people's seek of attention from others.

The literature on the economics of Peer-to-Peer(P2P) is also related to this paper since, broadly speaking, sharing resources on P2P networks is analogous to contributing content in online communities. Researchers have long been discussing the "free-riding" problem of P2P networks. Various incentive mechanisms have been proposed to tackle the "free-riding" problem [12, 13]. In a recent paper, Feldman et al. [14] developed a modeling framework that takes users' generosity into account. Although their paper focused very specifically on P2P networks, their suggestion that free-riding could be sustainable in equilibrium is very illuminating and reinforces our result on user segmentation in equilibrium. Also, the "generosity-driven" view is analogous to our "attention-driven" view. However, users' decision problem in their model is rather simplistic compared with our more comprehensive one. In this sense, our model significantly extends their work.

There is also a growing interest on user generated content in the computer science literature. For example, Huberman et al. [15] showed through an analysis of a massive data set from YouTube that the productivity exhibited in crowdsourcing, which is another name for user generated content, exhibits a strong positive dependence on attention, measured by the number of downloads. They found that lack of attention leads to a decrease in the number of videos uploaded and the consequent drop in productivity. In another recent paper, Guo et al. [16] empirically studied the patterns of user content generation in three online communities including a blog system, a social bookmark sharing network, and a question answering social network . They found that the user posting behavior in these three online social networks follows stretched exponential distribution, which is quite close to our finding of exponential [3] distribution in the empirical part. The major difference is that we derived the exponential distribution from our theoretical model with simplifying assumptions on the distribution of users' heteroteneity. The theoretical foundation of our empirical study differentiates our work from that literature.

[3] stretched exponential distribution with $c = 1$

3 The Model

3.1 Model Setup

There are n users in an online community where n is very large. In the current model, we assume that each user spends a fixed amount of time in the community which we call the time budget and denote by T_i for user i with $T_i \in [\underline{T}, \overline{T}], 0 < \underline{T} < \overline{T} < \infty$. There are two ways for a user to spend time in the community, consuming content or producing content. We use r_i to denote the proportion of time user i spends on reading content produced by other users and w_i the proportion of the time user i spends on producing content where $r_i + w_i = 1, r_i \geq 0, w_i \geq 0$. User i will produce $S_i = q_i T_i w_i$ amount of content where q_i is the productivity of user i in producing content, $q_i \in [\underline{q}, \overline{q}]$, $0 < \underline{q} < \overline{q} < \infty$. User i's decision variable is r_i or w_i.

Denote the total amount of content produced in the community by $S = \sum_{k=1}^{n} q_k T_k w_k$, and the total amount of content produced by everyone else except user i by $S_{-i} = \sum_{k=1,k\neq i}^{n} q_k T_k w_k$. Similarly, we denote $S_{-ij} = \sum_{k=1,k\neq i,j}^{n} q_k T_k w_k$.

A user obtains utility from two sources. First, since a user gets information or pleasure from the content, she obtains utility from consuming the content. The amount of utility a user can get by consuming content depends not only on the time he devotes to it but also on the amount of content available in the community. Hence we model this part of utility as follows

$$u_i^r = \psi(T_i r_i)\phi(S_{-i}) \tag{1}$$

where $\psi(0) = 0, \psi' > 0, \psi'' \leq 0$ and $\phi(0) = 0, \phi' > 0, \lim_{x \to \infty} \phi(x) < \infty$. The concavity assumption of ψ means decreasing marginal utility from consuming content. The monotonicity assumption of ϕ means that the more content available, the more utility a user can get from consumption for each unit of time. If there is no content available, then a user can't get any value from consuming so that the utility is zero ($\phi(0) = 0$). Although the increase of content available to user i will lead to increase of utility given $T_i r_i$, we believe such effect is bounded as S_{-i} goes to infinity so that $\lim_{x \to \infty} \phi(x) < \infty$. The rationale behind this is that a user can only get a limited amount of utility per unit of time regardless of how much content is available to her.

The second source for a user to obtain utility is by producing content. A user may enjoy publicity so that the more attention she gets from others who consume her content, the higher her utility. To model this, we assume that the utility a user gets from content production is proportional to the total attention she gets from all other users which we measure by the total time others spent on consuming her content. Intuitively, a user with more content should have a higher chance of getting attention from another user. To make things simple, we assume the amount of attention user i can get from user j is proportional to the content produced by i, which is

$$e_i^j = \frac{q_i T_i w_i}{S_{-j}} T_j r_j$$

In the above, the denominator is the total amount of content available to user j. So how much attention a user can get depends on her relative standing in terms of content in the community. Summing up over j gives the total attention user i gets, i.e.,

$$e_i = \sum_{j \neq i} \frac{q_i T_i w_i}{S_{-j}} T_j r_j$$

We use $\alpha_i \in [0, \overline{\alpha}]$ to account for users' heterogeneity in terms of their preferences for attention relative to that of content consumption. The larger α is, the more a user values attention. User i's utility function is then defined as follows

$$u_i = \psi(T_i r_i)\phi(S_{-i}) + \alpha_i \cdot \sum_{j \neq i} \frac{q_i T_i w_i}{S_{-j}} T_j r_j \tag{2}$$

Without taking into account the concavity of $\psi(\cdot)$, we could roughly interpret α as the value such that a user is indifferent between spending one unit of time consuming content and getting α unit of time from others reading her content.

The structure of the static game is described as follows. First, each user is endowed with (α, q, T) and the population endowment $(\alpha_i, q_i, T_i), i = 1, \cdots, n$ is common knowledge. Then, each user decides the proportion of time she will spend on content consumption (r_i) and content production (w_i) to maximize her utility. Although we start by assuming common knowledge of the population endowment $(\alpha_i, q_i, T_i), i = 1, \cdots, n$, later on, it turns out that in the limit equilibrium to be constructed, a user only needs to know her own endowment and the distribution of population endowment.

User's utility maximization problem could be formally written as

$$\max_{0 \leq w_i \leq 1} u_i = \psi(T_i(1 - w_i))\phi(S_{-i}) + \alpha_i \cdot \sum_{j \neq i} \left(1 - \frac{S_{-ij}}{S_{-ij} + q_i T_i w_i}\right) T_j r_j \tag{3}$$

which is a very complex problem in general since the solution depends on all other users' decisions. However, the decision problem is greatly simplified when the community size is large enough. The equilibrium concept we will use is Nash equilibrium and we are only interested in pure-strategy Nash equilibrium where each user chooses a $w_i \in [0,1]$. Although our results will hold under generic distribution of (α, q, T), for ease of illustration, we will assume from now on that the cumulative distribution function of $F(\alpha, q, T)$ is continuous in its support.

3.2 Homogeneous users

In the simplest case where users are homogeneous in the sense that $(\alpha_i, q_i, T_i) = (\alpha_0, q_0, T_0)$, we would expect each user of the comunity to serve both as a content producer and content consumer. Indeed, such kind of equilibrium always exists. Assume $w_j = w_0, r_j = 1 - w_0$, $j = 1, \cdots, i-1, i+1, \cdots, n$, then user i's utility function simplifies to

$$u_i = \psi(T_0(1-w_i))\phi\left((n-1)T_0 w_0 q_0\right) + \alpha_0(n-1)T_0(1-w_0)\frac{w_i}{(n-2)w_0 + w_i}$$

The first order condition is

$$-\psi'(T_0(1-w_i))\phi\left((n-1)T_0 w_0 q_0\right) + \alpha_0(n-1)(1-w_0)\frac{(n-2)w_0}{((n-2)w_0 + w_i)^2} = 0$$

We need to examine whether w_0 is the solution to the above equation. Substituting $w_i = w_0$ into the equation, one could easily prove that there always exists $w_0 \in (0,1)$ satisfying the first order condition. Therefore, in a homogeneous community, as long as the common α is positive, there exists a symmetric equilibrium where everyone spends time in both content consumption and content production.

3.3 Heterogeneous users

It would make more sense if we assume that community users are heterogeneous in terms of both productivity (q_i) and motivation (α_i). A natural question to ask then is how the consumption and production of content will be organized in such a heterogeneous community. For example, is there an equilibrium where everyone in the community serves as both content producer and content consumer as we have seen in a homogeneous community? Or on the other extreme, is there an equilibrium where users self-select themselves into either content consumers or producers but not both. In other words, we want to see if there is a trend of specialization among community users.

The community users could be classified into three groups in any equilibrium, those who only consume content, i.e., $I_C = \{i : w_i = 0\}$, those who only produce content, i.e., $I_P = \{i : w_i = 1\}$, and those who both consume and produce content $I_M = \{i : 0 < w_i < 1\}$. We call user i a consumer if $i \in I_C$, a producer if $i \in I_P$, and a prosumer if $i \in I_M$. Denote n_C, n_P, n_M to be the numbers of users in each group correspondingly with $n_C + n_P + n_M = n$. Also, we define $T^C = \sum_{i \in I_C} T_i$ which is the total amount of time spent by consumers, and denote $T^C_{-i} = T^C - T_i$ the total amount of time spent by all consumers except user i. The following lemma partially characterizes the equilibrium structure of a heterogeneous community.

Lemma 1. *In any equilibrium of the game, the three groups are characterized as follows*

$$
\begin{aligned}
i \in I_C &\iff \alpha_i q_i \leq h_C(i) = \frac{S_{-i}\phi(S_{-i})\psi'(T_i)}{T^C_{-i} + S_{-i}\sum_{j\in I_m, j\neq i} \frac{T_j r_j}{S_{-ij}}} \\
i \in I_P &\iff \alpha_i q_i \geq h_P(i) = \frac{(S_{-i} + q_i T_i)^2}{S_{-i}T^C_{-i} + (S_{-i} + q_i T_i)^2 \sum_{j\in I_M, j\neq i} \frac{S_{-ij}}{(S_{-ij}+q_i T_i)^2} T_j r_j}\phi(S_{-i})\psi'(0) \\
i \in I_M &\iff h_C(i) < \alpha_i q_i < h_P(i)
\end{aligned}
$$

Notice that the above lemma does not fully characterize each group as the RHS of each inequality depends on the decisions of each user in the community. However, it does suggest that the product αq measures to some extent the willingness of a user to produce content.

The next lemma is a technical result that we will use in Proposition 1. But the intuition is also not difficult to understand. Basically, it says that as the community size grows, the total attention (total consumption time) must increase at the same order with the amount of content which is at the order of total production time. In others words, the total consumption time and total production time must be balanced.

Lemma 2. *In any equilibrium* [4]

$$
\lim_{n\to\infty} \frac{T^C_{-i}}{S^\beta_{-i}} = 0, \quad \lim_{n\to\infty} \sum_{j\in I_M, j\neq i} \frac{T_j r_j}{S^\beta_{-ij}} = 0, \quad ,\forall i, j = 1, 2, \cdots, n, \forall \beta > 1
$$

Based on Lemma 1 and Lemma 2, we are able to characterize the asymptotic properties of $h_P(i)$ and $h_C(i)$, the thresholds that determine whether a user chooses to become a consumer, a producer, or a prosumer. The following proposition claims that there will be one common h_P for all users while $h_C(i)$ only depends on h_P and T_i as $n \to \infty$.

[4] We exclude one type of pathological equilibrium where $\lim_{n\to\infty} \frac{n_M}{n} \neq 0$, but $\lim_{n\to\infty} w_i = 0, \forall i \in I_M$

Proposition 1 (Segmentation). *The segmentation of the online community as its size becomes large enough is characterized by the following limit conditions:*

$$\lim_{n\to\infty} h_P(i) = h_P, \quad \lim_{n\to\infty} h_C(i) = h_P \frac{\psi'(T_i)}{\psi'(0)} = h_C(T_i),$$

so that all Nash equilibria of the game converges to the same limit equilibrium where

- *users with $\alpha_i q_i > h_P$ become producers,*
- *users with $\alpha_i q_i < h_P \frac{\psi'(\overline{T})}{\psi'(0)}$ become consumers,*
- *users with $h_P \frac{\psi'(\underline{T})}{\psi'(0)} < \alpha_i q_i < h_P$ become prosumers.*
- *those users with $h_P \frac{\psi'(\overline{T})}{\psi'(0)} < \alpha_i q_i < \frac{\psi'(\underline{T})}{\psi'(0)}$, depending on their T_i, they either become consumers or prosumers.*

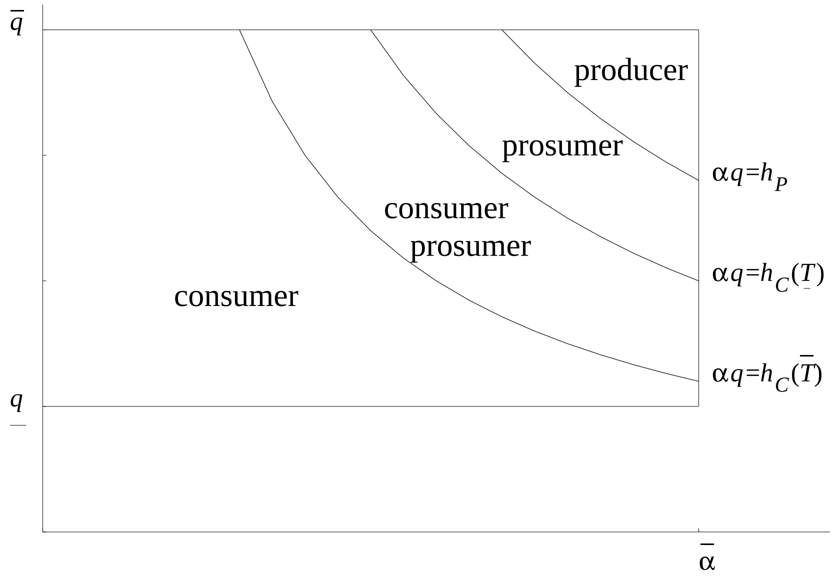

Figure 1: The Segmentation of the Online Community

The above proposition has several important implications. First, it suggests that when the community size is very large, each user's decision problem simplifies a lot. Rather than solving the complicated utility maximization problem of (3), she only needs to compare her $\alpha_i q_i$ with the two thresholds (h_P and $h_C(i)$) which are determined by the demand and supply of content in the equilibrium and her time budget. The thresholds, which could be roughly viewed as proportional to the ratio of content supply to content demand (attention), serve as the inverse of price for content in this particular online community. Each user compares her productivity and motivation with this inverse of price for content in the community to decide her role in the community. Second, Proposition (1) depicts how users in the community are divided into producers, consumers, or prosumers. Since $h_P \frac{\psi'(\overline{T})}{\psi'(0)} < h_C(T_i) < h_P \frac{\psi'(\underline{T})}{\psi'(0)}$, we have the segmentation of $\alpha - q$ plane shown in Figure 1.

Notice that the above result is in some sense similar to the result in a recent paper by Katona and Sarvary [17] where they studied the network structure of the commercial World Wide Web. They found that there is a specialization across sites in revenue models: high content sites tend to earn revenue from the sales of content while low content sites from the sales of traffic. In our case, highly motivated and more capable users obtain utility from content production while not highly motivated and less capable users obtain utility from content consumption.

From Lemma(2) and the proof of Lemma(1), we also have the following result which will be tested later in the empirical study.

Corollary 1. *If the community size is large enough, in any Nash equilibrium, (1)* $T_i =$

$T_k, q_i = q_k, \alpha_i > \alpha_k \Rightarrow w_i > w_k$ *(2)* $T_i = T_k, \alpha_i = \alpha_k, q_i > q_k \Rightarrow w_i > w_k$

The above result claims that in equilibrium, users with higher α or higher q devote larger proportion of time producing content, holding everything else equal, which is consistent with our intuition.

3.4 The Partition Equilibrium

Based on Proposition 1, when $\psi(r_i)$ is a linear function, $\lim_{n\to\infty} h_C(i) = lim_{n\to\infty} h_P(i) = h_P$. Under continuous distribution of (α, q), the proportion of those users with $h_C(i) < \alpha q < h_P(i)$ shrinks to zero as $h_C(i)$ and $h_P(i)$ become closer and closer. So we have the following result.

Corollary 2. *With continuous distribution of* (α, q, T), $\lim_{n\to\infty} \frac{n_M}{n} = 0$ *if and only if* $\psi'' = 0$, *i.e.,* $\psi(\cdot)$ *is linear.*

When $\psi(\cdot)$ is a linear function, we call the equilibrium coming out in the limit as $n \to \infty$ the partition equilibrium where the community is simply divided into two groups, content producers and content consumers. Such kind of equilibrium is of particular interest since it offers one possible explanation for the often observed phenomenon that in many online communities only a small proportion of people actually contribute content while the majority do not contribute anything. According to Proposition (2), in the partition equilibrium, there will be a clear division of labor. Those content producers will often be seen as very active users and those content consumers are often labeled as the "inactive" majority. Even though some may argue that those content consumers are free riders of the community, our model suggests that they are as important as those producers to sustain the community since the attention from them is the main drive for producers to keep contributing.

3.5 The Macro Dynamics of The Partition Equilibrium

Previously, we have taken T_i as given. Here, we assume that users have opportunity cost of time and allocate T_i to maximize the utility they can get from the community. For simplicity, we model this opportunity cost as a quadratic function $\frac{1}{2\theta_i}T_i^2$ where θ_i reflects the heterogeneity of opportunity costs among community users. We now extend the model to a two-stage game where users first choose the amount of time to spend in the community, and then choose w_i in the second stage.

The main purpose of this section is to show that the community supported by the partition equilibrium is robust in the sense that the macro-level production and consumption of content is stable even with perturbation of T_i in the first stage. This is in contrast with the instability of the contagious equilibrium discussed in [4]. We try to argue from a theoretical point of view that the vanity-based or attention-driven perspective could be a more practical explanation to the thriving of user-generated contents. People contribute content to the community not because they fear the community will collapse as a result of them not contributing, but because they obtain utility from getting attention.

Assume that $\psi(\cdot)$ is linear, and denote $\psi'(T) = \tau > 0, \forall T \geq 0$ for convenience. Also, we normalize $\lim_{S\to\infty} \phi(S) = \frac{1}{\tau}$ so that $\psi'(T) \lim_{S\to\infty} \phi(S) = 1, \forall T > 0$. This is without loss of generality since we could scale α for the whole community to adjust $\phi(S)$. We also assume

n is very large. These assumptions ensure that users in the first stage expect the partition equilibrium will be played in the second stage and that analytical results on the stability of the community could be obtained.

First, we characterize the equilibrium in the two-stage game.

Proposition 2. *The subgame perfect equilibrium of the two-stage game is characterized as followers: (1) User with $\alpha_i q_i < h = \frac{S}{T^C}$ chooses $T_i = \theta_i \phi(S)\tau$ in the first stage and $r_i = 1, w_i = 0$ in the second stage; (2) User with $\alpha_i q_i > h = \frac{S}{T^C}$ chooses $T_i = \frac{T^C}{S}\theta_i \alpha_i q_i$ in the first stage and $r_i = 0, w_i = 1$ in the second stage. S and T^C are determined as the solution to the following equations*

$$\begin{cases} S = \sum_{\alpha_i q_i > \frac{S}{T^C}} \frac{T^C}{S} \alpha_i q_i^2 \theta_i \\ T^C = \sum_{\alpha_i q_i < \frac{S}{T^C}} \phi(S)\tau\theta_i \end{cases} \quad or \quad \begin{cases} S = \int_{\alpha q > \frac{S}{T^C}} \frac{T^C}{S} \alpha q^2 \theta dF(\alpha, q, \theta) \\ T^C = \int_{\alpha q < \phi(S)} \tau\theta dF(\alpha, q, \theta) \end{cases} \tag{4}$$

where $F(\alpha, q, \theta)$ is the distribution of users in the community.

The pair of equations (4) implicitly gives the macro-level supply and consumption of content in the equilibrium. It also defines a time-invariant dynamic system of content production and consumption at the macro-level. Due to page limit, we do not include the definitions for dynamic system, equilibrium point, and asymptotically stable equilibrium point here. Readers can see ([18]) for reference. The next proposition claims that (S, T^C) determined in the equilibrium identified in Proposition (2) is an asymptotically stable equilibrium.

Proposition 3 (Macro-level Stability)**.** *The macro-level dynamics of content production and* *consumption is characterized by Equations (5)*

$$\begin{cases} T^C = g_1(S) = k_1\phi(S) \\ \\ S = g_2(T^C) = k_2\sqrt{T^C} \end{cases} \tag{5}$$

where $k_1 = \sum_{i\in I_C}\theta_i$ *and* $k_2 = \sqrt{\sum_{j\in I_P}\alpha_i q_i^2\theta_i}$ *are constant. This dynamic system has an asymptotically stable equilibrium point* (T^{C*}, S^*) *where* $T^{C*} > 0$, $S^* > 0$.

Figure 2 shows an example of the dynamic system. From the figure, we could roughly see the stability of this system. Notice that $(T^C, S) = (0, 0)$ is also an asymptotically stable equilibrium point. Hence, none of the two equilibria are globally stable.

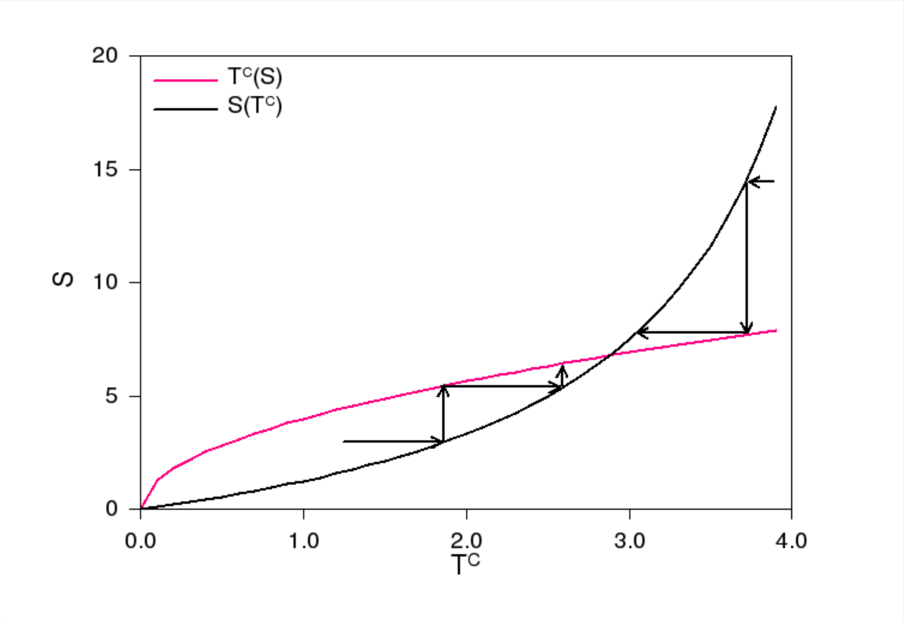

Figure 2: Equilibrium of content supply and demand

4 Empirical Test

4.1 Data Description

We collected data from Twitter [5], which is an open social networking and micro-blogging service launched publicly in July, 2006. It is probably the fastest-growing social network site in 2009 and is estimated to have tens of millions users. Users can use Twitter to post and read messages known as tweets which are text-based posts of up to 140 characters. A user's followers are the people who subscribe to receive the user's tweets. Users don't necessarily need to know the people they are following and vice versa. Twitter has an open API (Application Programming Interface) which we have used to develop a program to collect user information including user name, location, number of updates, number of followers, number of people they are following (called friends in the API documentation), and a short description of him/herself.

Our data comprises of 3.26 million Twitter user profiles collected in July, 2009, among which we use about 1.3 million profiles for model estimation. The total number of Twitter users is estimated around 30 million. For our research, we are mainly interested in the frequency of posting new tweets or updates by each user. To obtain this, we divide the total number of updates by the total number of days since an account was created which are both available in our data. We denote the derived variable by $UPDATERATE_i$, which we believe is a reasonable proxy of $T_i w_i$ in our model. This is because the length of a tweet is limited to 140 characters so that on average the time needed to write a tweet should not

[5] http://twitter.com

differ much. However, the amount of content in each tweet depends on the capability of the producer, i.e., q_i. Some content producers post very informative or insightful tweets while some producers post mostly trivial or even spam tweets.

Users on Twitter have the option of putting a URL on their profiles. This URL will be displayed on their Twitter homepages and can be clicked by visitors who might be interested in this particular user. We use a dummy variable URL to capture this difference. We also include the number of followers each user has and the number of friends each user has. [6] Table 1 summarizes the variables used in the empirical study.

Table 1: Summary of Variables

Name	Meaning	Min	Max	Mean
$UPDATES$	number of tweets posted	0	13065	146.93
URL	1 if user has URL in profile, 0 otherwise	0	1	0.1961
$FRIENDS$	number of friends	0	77445	386.82
$FOLLOWERS$	number of followers	0	1994900	416.81
$DAYS$	number of days since account creation	5	1096	62.071
$UPDATERATE$	$UPDATES/DAYS$	0	0.2	15

[6] These two number may contain a lot of noise due to the "following back" etiquette, i.e., a user usually follows back another user who follows him.

4.2 Hypotheses

About 20% of the Twitter users put a URL on their profiles while the other 80% don't. The URL users put on their profiles usually link to their personal blog sites or the organization they represent. These are generally sites they try to promote to attract visitors. We believe users with URL on their profiles value attention more than those without URL on their profiles for two reasons. First, the fact that some users put URL on their profiles while others do not is the result of user self-selection based on how much they value attention. For those who highly value attention, they are more likely to put URL on their profiles to actively seek attention in the form of clicks on their URL. On the other hand, those users who do not care so much about attention will not bother to set up a personal or organizational page and put that on their profiles. This underlying self-selection process leads to the result that users with URL on their profiles value attention more than those without URL on their profiles. Secondly, now that some users have a URL on their profiles, each visit by another user brings additional value due to potential clicks on their URL, hence making those users with a URL value attention even more. While both factors might be at work, we believe the underlying self-selection process is more important.

Translating into our model, the above arguments suggest that those with URL on their profiles have larger α values than those who don't. On the other hand, Corollary(1) implies that users with larger α values will devote larger proportion of time producing content in equilibrium. Assuming α, q, T are distributed independently, we have the following hypothesis,

Hypothesis 1. *Twitter users with URL on their profiles will have higher* $UPDATERATE$ *than those users without URL on their profiles.*

The number of followers a Twitter user has is an indicator of a user's capability of producing content, although it may contain a lot of noise. Users often follow others who post tweets interesting to them, Analogous to the idea of PageRank, we would expect users with more followers to be more productive in the sense that they are more capable of producing content that others find interesting or useful. So we take $FOLLOWERS$ as a proxy of q in our model. Based on Corollary(1), and again by assuming that α, q, T are distributed independently, we propose the following hypothesis

Hypothesis 2. *Twitter users with higher* $FOLLOWERS$ *have higher* $UPDATERATE$

To test the above hypotheses, we use the following regression model

$$\ln(UPDATERATE_i) = b_0 + b_1 \times URL_i + b_2 \times \ln(FOLLOWERS_i) + b_3 \times \ln(FRIENDS_i) + \epsilon_i \tag{6}$$

If our hypotheses is correct, the estimates of b_1 and b_2 from Equation 6 should be positive and significant.

Our third hypothesis regards the distribution of $UPDATERATE$ among Twitter users. Proposition (1) suggests that as the community size becomes large enough, it will be partitioned into three groups where users in the bottom left region of the $\alpha - q$ plane become content consumers, users in the upper right region become content producers, and users in between these regions become prosumers (or consumers), as is shown in Figure (1). So we would expect a large proportion of Twitter users with very low $UPDATERATE$.

To derive the distribution of $UPDATERATE$, we again use the setup of the two-stage-game we used in Section 4.3. except that we don't assume $\psi''(\cdot) = 0$ here so that we would expect three segments of the community, i.e., consumers, producers, and prosumers. It's easy to deduce that $T_i = \alpha_i q_i \theta_i \sum_{j \neq i} \frac{T_j r_j}{S_{-j}}, i \in I_P$. Hence producers with the same $\alpha q \theta$ value should spend the same amount of time producing content, hence the same $UPDATERATE$. We will need assumptions on the distribution of (α, q, θ) to derive the distribution of $UPDATERATE$. For simplicity, we assume that the distribution of θ_i is independent of that of (α_i, q_i) and (α_i, q_i) is uniformly distributed in the region $(0, \overline{\alpha}) \times (\underline{q}, \overline{q})$. Hence, the number of content producers who spend time T, $T_2 < T < T_1$ producing content is proportional to the shadow area in Figure (3), which, denoted by $A(T_1, T_2)$, is:

$$A(T_1, T_2) = (T_2 - T_1)(1 + \ln\bar{\alpha} + \ln\bar{q}) + T_1 \ln T_1 - T_2 \ln T_2$$

$$\lim_{T_2 \to T_1} \frac{A(T_1, T_2)}{T_2 - T_1} = ln\bar{\alpha} + ln\bar{q} - lnT_1$$

So the following econometric model should characterize the distribution of $UPDATERATE$ among producers, which accounts for those larger $UPDATERATE$ values.

$$USERCOUNT_k = b_0 + b_1 ln(UPDATERATE_k) \tag{7}$$

where $USERCOUNT_k$ is the number of user whose $UPDATERATE$ is in the interval

$$[UPDATERATE_k, UPDATERATE_{k+1}]$$

We use the following cutoff points to categorize all the Twitter user in our sample.

$$UPDATERATE_0 = 0, UPDATERATE_{k+1} = UPDATERATE_k + step, step = 0.01, k = 0, 1, \cdots, 499$$

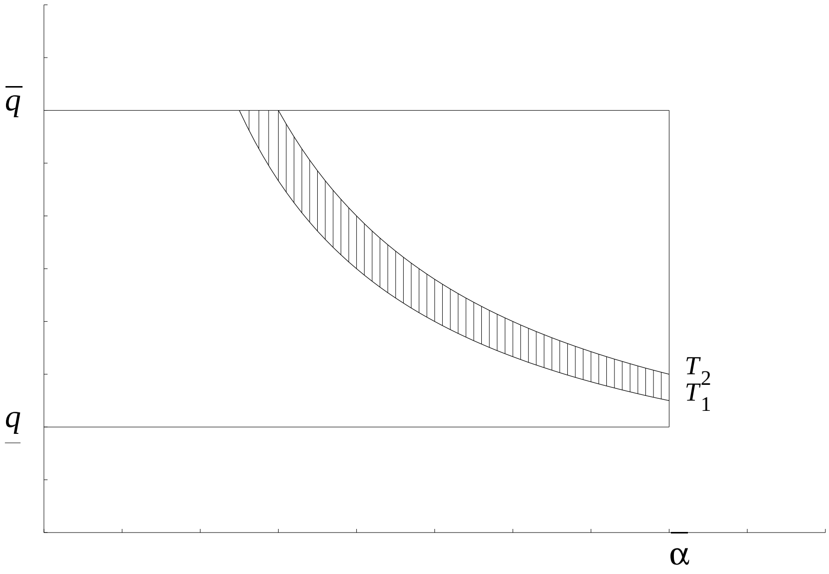

Figure 3: Hypothesis 2

On the other hand, the above econometric model will not fully capture the distribution of $UPDATERATE$ of prosumers whose amount of time spent on production generally depends not just on the product of αq but also on the values of α, q separately.

Based on the above analysis, we conjecture that Equation (7) should fit the data better when we exclude small values of $UPDATERATE$, and this effect should be most evident when we exclude the first point $UPDATERATE_0$ which accounts for the large proportion of consumers and will diminish as we exclude $UPDATERATE_i, i = 1, 2 \cdots$ since prosumers should behave more and more like producers as their αq values get close to h_P.

For ease of illustration, we denote this new sample with all 500 points as $SAMPLE_0$, the subsample of $(USERNUM_k, UPDATERATE_{k+1}), k = 1, 2, \cdots, 499$ as $SAMPLE_1$, and so on. We also denote R_m^2 as the R-squared of the regression of Equation (7) with $SAMPLE_m$. Figure (4) shows the first 50 points of $SAMPLE_0$ in the form of a histogram. The horizontal axis is $UPDATERATE \times 100$ and the vertical axis is the number of users. So each bin represents the users with $UPDATERATE$ falling in the range of the bin.

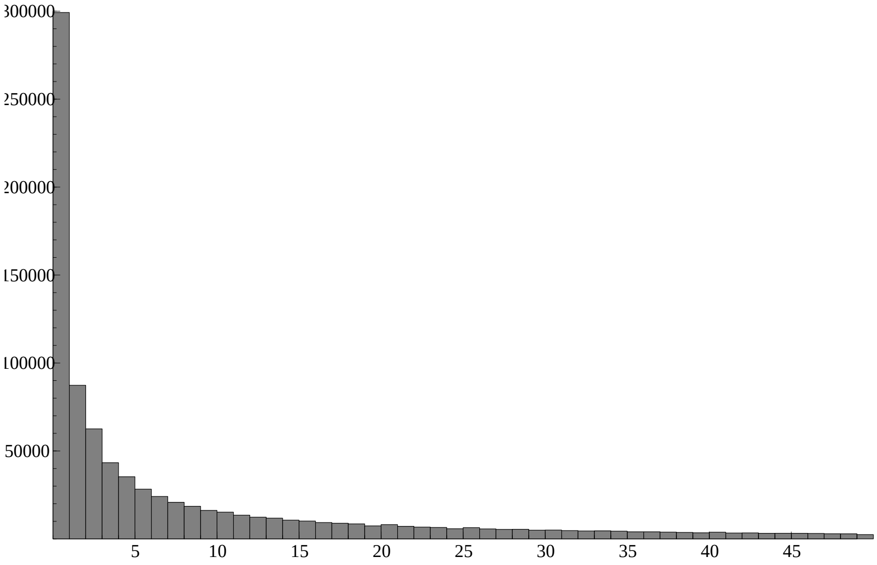

Figure 4: Histogram of Updates by sample Twitter user

We consider R-squared as a reasonable measure of how well Equation (7) fits the data. Hence, we have the following hypothesis.

Hypothesis 3. *(H3a) R_1^2 is significantly larger than R_0^2*

(H3b) When m is small, R_m^2 increases with m but the increment becomes smaller as m increases. In other words, R_m^2 should be an increasing and concave function of m when m is not too large.

4.3 Results

We have collected a total number of 3,261,218 Twitter user profiles, among them 1,521,136 joined Twitter at least 90 days before we collected their profiles. We impose this 90 days restriction to get reasonably accurate estimates $UPDATERATE$. Among these Twitter users, we further remove those who have 0 friends, or 0 followers, or 0 updates so that we could do the logarithmic transformation on the data. We finally obtain 1,304,117 valid user profiles.

Table 2 shows the estimation results of Equation (6). Both the coefficient of the dummy variable URL and $\ln(FOLLOWERS)$ are significantly positive, which supports (H1) and (H2). We also notice that $\ln(FRIENDS)$ is also significantly positive although the effect is much weaker compared with URL and $\ln(FOLLOWERS)$. This is reasonable since Twitter users sometimes engage in conversation with friends which would lead to more updates.

We then use the same sample of Twitter users to construct $SAMPLE_k, k = 0, 1, \cdots, 99$ and estimate Equation (7). As a comparison, we also use 100 randomly generated subsamples from $SAMPLE_0$ with size from 400 to 499 [7] to estimate Equation (7).

In all these regressions, the coefficient b_1 is negative and significant. Figure (5) shows how the R-squared changes as we use different subsamples to estimate. As we can see, for the randomly selected subsamples, there is no clear relationship between the R-squared and the size. However, for the R-squared of $SAMPLE_k, k = 0, 1, \cdots, 99$, the trend is very evident. First, by switching from $SAMPLE_0$ to $SAMPLE_1$, the R-squared increases from 0.266

[7] The first point is excluded

Table 2: OLS Estimation Results

Variable	OLS Estimates
URL	0.44448*** (128.04)
$\ln(FOLLOWERS)$	0.83032*** (460.38)
$\ln(FRIENDS)$	0.0092030*** (4.7371)
$Constant$	-4.2946*** (1066.1)
R^2	0.49053
Observations	1,304,117

t-values are in parentheses.

*10% significance, **5% significance, ***1% significance

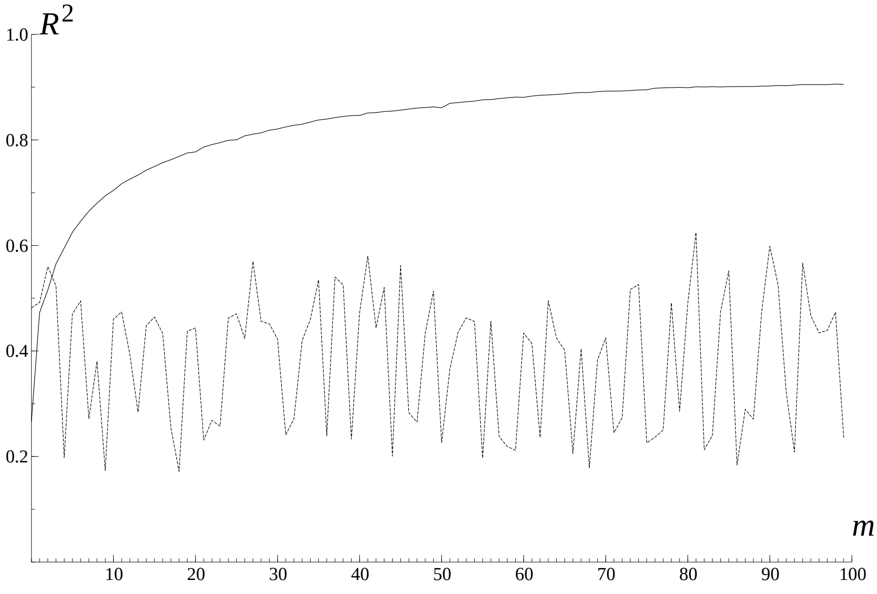

Figure 5: R^2 of $SAMPLE_m$

to 0.473, which results from the fact that those content consumers are excluded from the samples. The R-squared increases rapidly as we move from $SAMPLE_1$ to $SAMPLE_{20}$ which is caused by the gradual elimination of prosumers from the samples. The R-squared keeps increasing as we increase m but at a much slower pace. This is reasonable since prosumers removed during this stage behave more like producers hence are better characterized by Equation (7). As m reaches around 80, however, the R-squared stabilizes around 0.9. It means that users removed from now on are mostly producers and further removal will not improve the model fitness anymore. The concavity of the curve when m is smaller than 80 is clear in the figure. Based on the above analysis, we believe both H3a and H3b are well supported by the data.

5 Conclusion and Limitation

We started with the question of why users contribute content in online communities and how their incentive structure affects the collective outcome of user content generation in online communities. As in most literature, we have assumed that users derive utility from consuming content. However, unlike the traditional economic literature which model user content generation as a purely cooperative behavior, we have assumed that users actually obtain utility from content production by getting attention from others. Although the underlying mechanisms of how users obtain utility from attention might vary from economic incentives to psychological and sociological motivation, we treated them abstractly as users' taste for attention. Combining these two sources of utility for users in online communities, we derived a very general utility function based on which we constructed a stage game to characterize the interaction among users. We take into account user heterogeneity in the utility function through two individual parameters, α and q, which capture users' individual taste for attention and the productivity of generating content.

To study the outcome of the game, we use the concept of Nash equilibrium and rely on the assumption that the size of the online community is very large. We find that as the size of the online community goes to infinity, all Nash equilibria of the game converges to the same limit equilibrium where users choose their roles in the community solely based on the value of their endowment (T_i) and the product of α and q. Roughly speaking, users with very high αq values tend to become content producers who only produce content without consuming any content, and users with very low αq values tend to become content consumers who only consume content without producing any. The proportion of prosumers, who both produce and consume content, crucially depends on how concave the consumption part of the utility function $\psi(\cdot)$ is. The more concave $\psi(\cdot)$ is, which means the less marginal utility a user will get if she spends more time consuming content, the larger the prosumers' proportion is. In the extreme when $\psi(\cdot)$ is linear, there will be no prosumers in the online community. Users either become content producers or content consumers. We have showen that under the assumption of partition equilibrium, the macro-level content consumption and production, characterized by a time-invariant dynamic system, is stable in the sense that there is an asymptotically stable equilibrium point involving massive content production and consumption. This result gives strong support to the sustainability of user content generation in online communities.

To bring theory to practice, we investigated the distribution of content contribution of over 1 million Twitter user profiles. Overall, the empirical study supports our model quite well.

Although we believe our work has provided some important insights into the phenomenon of user content generation, we realize that it is limited in both theoretical modeling and empirical study. First of all, we haven't fully characterized the limit equilibrium of the general case when $\psi(\cdot)$ is nonlinear since the model will soon become intractable when we try to solve analytically each prosumer's optimal w. One way to extend our model is to run computer simulation to further explore properties of the limit equilibrium. Our empirical study is also limited due to our lack of data. Since both α and q are unobservable, we have to find proxy variables to indirectly test our model. In our current dataset, URL is a very coarse proxy for α and $FOLLOWERS$ also contains a lot of noise. Finding a better dataset from Twitter or other large online communities is an obvious direction for future research. Another future research direction is to extend this paper to dynamics, both theoretically and empirically. Users may start out with no information at all about an online community. How do they adjust their behavior dynamically as they learn more about the online community? It would be interesting to capture this through modeling as well as through empirical studying.

User content generation is now pervasive on the Internet. The popularity of Youtube, Flickr and Twitter is a fascinating glimpse of its huge impact on our economy and society. This paper is only a small attempt to understand the drive and structure of user content generation. There are tremendous opportunities for both theoretical research and empirical study to better understand the structure and evolution of user content generation.