# Analyzing Patterns of User Content Generation in Online Social Networks

Lei Guo[1], Enhua Tan[2], Songqing Chen[3], Xiaodong Zhang[2], and Yihong (Eric) Zhao[1]

[1]Yahoo! Inc.
701 First Avenue
Sunnyvale, CA 94089, USA
{lguo,yzhao}@yahoo-inc.com

[2]Dept. of Comp. Sci. & Engr.
Ohio State University
Columbus, OH 43210, USA
{etan,zhang}@cse.ohio-state.edu

[3]Dept. of Computer Science
George Mason University
Fairfax, VA 22030, USA
sqchen@cs.gmu.edu

## ABSTRACT

Various online social networks (OSNs) have been developed rapidly on the Internet. Researchers have analyzed different properties of such OSNs, mainly focusing on the formation and evolution of the networks as well as the information propagation over the networks. In knowledge-sharing OSNs, such as blogs and question answering systems, issues on how users participate in the network and how users "generate/contribute" knowledge are vital to the sustained and healthy growth of the networks. However, related discussions have not been reported in the research literature.

In this work, we empirically study workloads from three popular knowledge-sharing OSNs, including a blog system, a social bookmark sharing network, and a question answering social network to examine these properties. Our analysis consistently shows that (1) users' posting behavior in these networks exhibits strong daily and weekly patterns, but the user active time in these OSNs does not follow exponential distributions; (2) the user posting behavior in these OSNs follows stretched exponential distributions instead of power-law distributions, indicating the influence of a small number of core users cannot dominate the network; (3) the distributions of user contributions on high-quality and effort-consuming contents in these OSNs have smaller stretch factors for the stretched exponential distribution. Our study provides insights into user activity patterns and lays out an analytical foundation for further understanding various properties of these OSNs.

## Categories and Subject Descriptors

H.1 [**Information Systems**]: Models and principles

## General Terms

Human Factors, Measurement

## Keywords

User generated content (UGC), social networks, stretched exponential distribution

## 1. INTRODUCTION

Recent years have witnessed the success of a number of online social networks (OSNs), such as Del.icio.us (`http://delicious.com/`), Facebook (`http://www.facebook.com/`), Flickr (`http://www.flickr.com/`), LinkedIn (`http://www.linkedin.com/`), Yahoo! Answers (`http://answers.yahoo.com`), and YouTube (`http://www.youtube.com/`). These social networks have attracted a significant number of participants that contribute various contents on the Internet, which is often referred to as user generated content (UGC), owing to the pervasive broadband Internet accesses and the ever-increasing bandwidth available to end users [4].

Users are basic elements of these OSNs and communities. In general, a user's activities in OSNs include authoring content, viewing, and networking. According to their different purposes, existing OSNs can be classified into two categories, the *networking oriented* OSNs and the *knowledge-sharing oriented* OSNs. The former, such as Facebook and LinkedIn, emphasizes more on the networking perspective, and the social relationship is the basis of these OSNs. Hence, we call them *networking oriented* OSNs. In these OSNs, content sharing is mainly among friends. The latter, such as blog networks, question answering networks, and viral video networks, emphasizes more on the knowledge or content sharing. Thus, we call them *knowledge-sharing oriented* OSNs. The network in these OSNs is not driven by the underlying social relationships. Instead, the network is formed through the users' common interests on the shared content.

The rapid development of these OSNs has attracted significant attentions from research community. A number of studies [1, 2, 5, 13, 16] have been conducted to examine various properties of different OSNs. For example, Cheng et al. [5] studied the YouTube videos and found that the links to related videos generated by uploaders' choices have clear small-world characteristics, which indicates that the videos have strong correlations with each other. In work [16], based on four large online social networks, the authors studied the evolution of social networks and showed that the combination of the gap distribution with the node lifetime leads to a power law out-degree distribution that accurately reflects the real network in all four cases. In blogspace, works [1, 2] have studied the link propagation and information epidemics.

However, these existing studies mainly focused on how users are connected and thus how the networks are formed, as well as how a social network graph evolves over time, such as [16]. Users who have a large number of connections are the core of social networks, and play an important role on information propagation. But in knowledge-sharing oriented OSNs, how users participate in the network and how users generate and share content play the key role in attracting viewers, since the user participation and contribution in these OSNs drive the growth of these social network communities and the success of their business. Therefore, understanding the patterns of user participation and user posting behavior in these knowledge-sharing oriented OSNs is very imperative to social network industry and researchers, in order to identify and distinguish

*KDD'09,* June 28–July 1, 2009, Paris, France.
Copyright 2009 ACM 978-1-60558-495-9/09/06 ...$5.00.

active users from spamming users, attract new users and keep existing users, predict hot spots and the trends of topics in user communities, and perform efficient resource management in the underlying supporting system.

The user participation in terms of active time in peer-to-peer and social networks has been assumed to follow exponential distributions in modeling [9, 16], and it has been also assumed that there is strong correlation between user active time and user contribution. It had been reported that in Wikipedia, most of articles are contributed by a small number of users [24]. In [23], Voss found that both the number of Wikipedia articles a user edited and the number of authors for a Wikipedia article follow power law, for contents of different languages. In [12], Kittur et al. analyzed the edit logs of Wikipedia from 2001 to 2006, and find a shift of the user contributions from a core group of elite users to common users, in terms of both number of edits and length of revised contents. A similar shift of user contribution distribution for Del.icio.us social network was also reported in [12]. However, whether the user contribution in Wikipedia/Del.icio.us follows power law or not is not further analyzed. On the other hand, some studies have been conducted on the distribution of user participations in networking oriented OSNs. For example, Seshadri et al. studied the mobile phone call graph as a social network and analyzed the distribution of the length of mobile calls [21]. They found it does not follow power law or lognormal. Instead, the double Pareto LogNormal fits the data very well. Gjoka et al. [7] studied the applications on Facebook and found that although the number of application installations increases with time, the average user activity decreases. These findings have put the commonly accepted power law assumption in doubt.

In this work, we set to study the user contributions and activities in knowledge-sharing oriented OSNs empirically. For this purpose, we have analyzed three workloads of popular OSNs, including blog, bookmark sharing, and question answering social networks for a duration of multiple years. In these OSNs, we are particularly interested in how users participate in the network, generate or post content, as well as the quality of the content. We have the following findings in this study:

1. User posting behavior of original content in these OSNs shows strong daily and weekly patterns. However, for non-original content posting (i.e., content cut-and-pasted from other sources), we have not observed temporal posting patterns (or patterns along time).

2. We have observed two groups of users with distinct posting behaviors: (1) steadily posting in the network, (2) inactively posting. The rest users post occasionally in the network. The overall user lifetime does not follow the exponential distribution.

3. The distribution of different users' posting activity does not follow power law distributions. Our analysis across three workloads from both short terms (in weeks) and long terms (in years) consistently shows that it follows the stretched exponential distribution, for which the individual contributions of top users are distributed much flatter than those in power law networks.

4. We have observed that the stretched exponential distribution of user contributions in OSNs roughly follows the "80-20" rule, i.e. 20% users contribute 80% total content in the network. However, the cumulative contribution ratios of a small number of top-$k$ users in the OSNs are much smaller in OSNs than those in standard power law distributions, such as Pareto's social wealth distribution. This implies that contents in an OSN are not mainly contributed by a small number of core users.

5. Users contribute different types of UGC objects. Their contributions can be characterized by the stretched exponential model with different parameters. For example, the distribution of user contributions on high-quality content tends to have a small stretch factor, indicating that it is more skewed towards (a few) core users.

Our findings provide insights into user behaviors of OSNs, which is timely desirable in social network industry and research community. We hope our analysis lays out a foundation to guide the design, modeling, and simulation of large and complex OSNs with various properties for many applications.

The remainder of this paper is organized as follows. We present related works in Section 2. In Section 3, we overview the social network systems studied in this work. Our detailed workload analysis is conducted in Section 4. We further analyze the implications of our findings in Section 5 and make concluding remarks in Section 6.

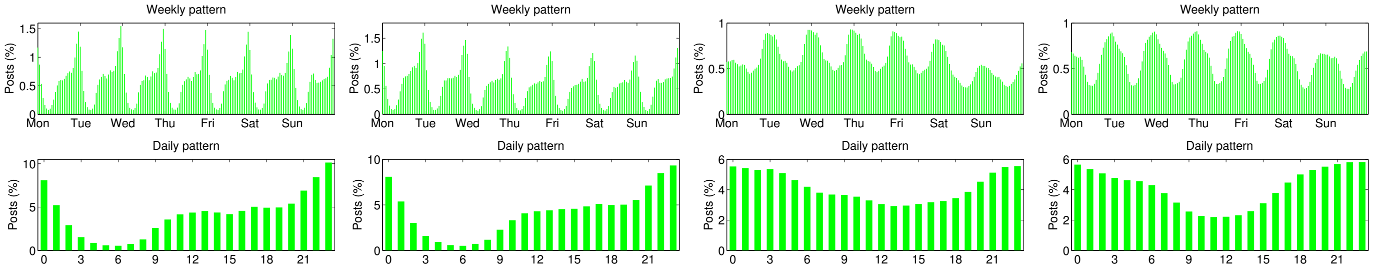
## 2. RELATED WORK

Online social networks have attracted considerable attention recently. A number of studies have been conducted on different forms of social networks. For example, as one of the typical social network domain on the Internet, blogspace has constantly attracted researchers' attention. Earlier work [1, 2, 13] had focused on the link propagation behaviors in blogsphere and studied the information epidemics. While cascading behaviors were rare in [2, 8], Leskovec et al. [19] analyzed about 45 thousands blogs and 2.2 million postings, reported that the size of the cascades distribution follows a perfect Zipf distribution with the slope value of $-2$. Accordingly, a flu-like epidemiological model is proposed to characterize the cascades in blogspace. Gruhl et al. [8] studied the dynamics of information propagation at both topic and individual levels in blogspace. They showed that topics are composed as a union of long term "chatter" topics and short-term "spike" topics. At the individual level, they propose a model for information diffusion based on the spread of infectious diseases [3].

Besides blogs, various other online social networks have also been studied. For example, Guo et al. [27] studied the instant messaging (IM) networks and found that the social network of IM users does not follow a power law distribution; instead, it can be characterized by a Weibull distribution. Leskovec et al. [17] studied Microsoft Messengers and found the average path length among users is 6.6.

The empirical results of social network community structure and evolution have also been reported. For example, Leskovec [18] et al. analyzed about 70 large networks and found that community structure is different from what have been reported, and proposed a "forest fire" generative model to characterize such structures. In work [16], based on four large online social networks, authors studied the evolution of social networks and showed that the combination of the gap distribution with the node lifetime leads to a power law out-degree distribution that accurately reflects the true network in all four cases.

On the other hand, analytical models have also been studied for social networks for a long time. For example, work [20] aimed to find the most influential nodes and build probabilistic models for viral marketing, and work [11] tried to find the most influential nodes in several of the most widely studied models in social network analysis. In [26], authors performed theoretical analysis on information cascades based on random graphs while in [15] authors analyzed topological cascade patterns in a large product recommendation network empirically.



(a) Blog Article posting (b) Blog Photo posting (c) Bookmark posting (d) Answer posting

**Figure 1: Weekly and daily patterns of UGC posting in OSNs**

Among these social networks, user activities and the UGC play a key role, particular to knowledge-sharing oriented OSNs. Work [25] proposed fast algorithms to characterize bursty patterns of user posting activities in blogspace. Such bursty patterns are often regarded as the result of heavy-tailed dynamics, and power law distributions [6, 23] are often used to characterize such attributes. In P2P networks, Stutzbach and Rejaie have also studied the churns caused by user activities [22].

Since user activity patterns in knowledge-sharing oriented OSNs have not been well studied despite of their significance, we set to study such patterns and its implications in this work. Our empirical analysis provides several new findings that are different from the commonly accepted concepts about user activities and contributions in these OSNs.

# 3. USER ACTIVITY AND UGC CONTENT CREATION OVER TIME

In this study, we analyze three OSN workloads. The *Blog* workload contains the DB dump of a blog system with millions of users in Asia, including all posts of articles and photos for multiple years. We will refer them as *Blog Article* (or *Article*) and *Blog Photo* (or *Photo*), respectively. The *Bookmark* workload is from the DB dump of a well known social bookmark sharing network in U.S., including all bookmarks posted for a number of years. The *Answer* workload contains the DB dump of a famous question answering social network in U.S., including all question and answer posts for several years. In our study, we only consider objects posted by users. User comments, such as comments of a blog article by the blogger's friends, are not considered as UGC posts. Before we present the detailed analysis results, we first present an overview of user posting activities and the posted content in this section.

## 3.1 Daily and weekly patterns

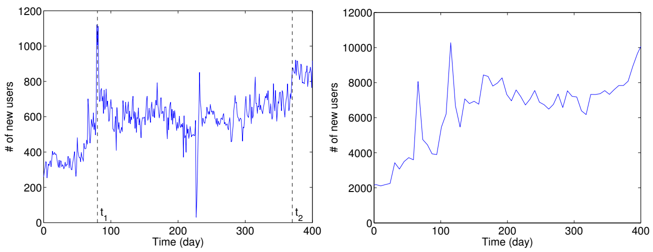
All three social networks in our study are expanding with time. In general, the number of daily new posts increases with time sub-linearly, while the cumulative number of all posts increases with time super-linearly. For Bookmark and Answer, the daily number of new posts for weekdays and weekends show different patterns. To understand these weekly patterns further, we study the number of posts in different weekdays and weekends. We bin UGC posts by hours, and count the numbers of posts in the same hours by weeks, for the three OSNs. We then normalize the number of posts in each hour by the total number of posts across the entire trace duration. The upper plots of Figure 1 show the normalized weekly posts for Blog Article, Blog Photo, Bookmark, and Answer, respectively. We can see for Bookmark and Answer systems, the hourly number of posts in weekdays is much higher than the hourly number of posts in weekends. For Blog systems, the hourly number of posts is similar during weekdays and weekends. This is because blogging is a kind of daily Web journaling or diary writing, so the daily user activities on blog posting do not change dramatically across different days in a week.

The bottom plots of Figure 1 further show the daily patterns of UGC posting, computed in a similar way as that of weekly patterns of UGC posting. For all three OSNs, the peak time for posting is about 23:00 in local time. However, the least active time for blog article/photo posting is 6:00, which is different from bookmark/answer posting (12:00-13:00), indicating different user activity patterns in these OSNs.

## 3.2 User/content increase rate in OSNs

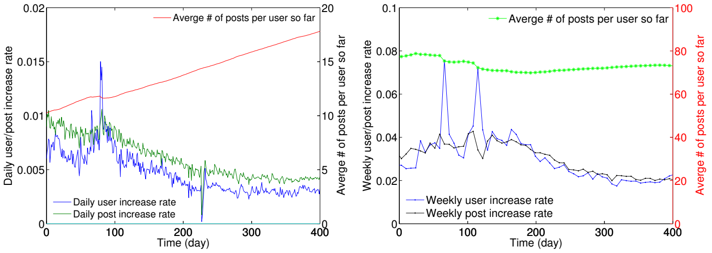
To study the statistics of user contributions in OSNs, a common method is to consider the union of all users and all posts in the system during a measurement period. However, users may join the system at different time in this period. For large scale measurements, the number of new users in this period can be non-trivial and the joining rate can be bursty. Figure 2 shows the daily number of new users joined in Blog and Bookmark systems for a selected period in the workload duration (the starting dates are hided for the purpose of business confidential protection). In addition to the weekly and daily fluctuating patterns of newly joined users, the user joining events are bursty in daily, weekly, and even a larger time scale. For example, as shown in Figure 2(a), around time $t_1$, in the trace collection duration, the number of new users is significantly higher than that before and after $t_1$. For Bookmark, as shown in Figure 2(b), there are a number of spikes, indicating bursts of user joining events spreading over a number of weeks.

Figure 3(a) shows the *daily* increase rate of users and article posts on the left $y$-axis, and the overall contribution per user on the right $y$-axis for the Blog system with time, for the same time period in Figure 2(a). The daily user (post) increase rate is defined as the ratio of the number of newly joined users (newly posted articles) in a day over the cumulative number of users (articles) by the last day. The overall contribution per user by a day is defined as the ratio of the cumulative number of posts in the system by that day over the cumulative number of users that have joined the system by that day. In general both the post increase rate and the user increase rate decrease with time gradually, meaning the total numbers of both posts and users increase with time super-linearly but not exponentially. Furthermore, the post increase rate is always greater than the user increase rate except around $t_1$. We can see that both the daily user increase rate and the post increase rate around $t_1$ are greater than those in the neighboring days noticeably. However, the ratio of the post increase rate to the user increase rate in that day is smaller than those in the neighboring days. This means the burst of these new users have less contributions than the average user in the system. As a result, the overall contribution per user does not increase on that day and the increase becomes linear after that. In contrast, another user increase burst around $t_2$ does not affect the overall contribution per user much.



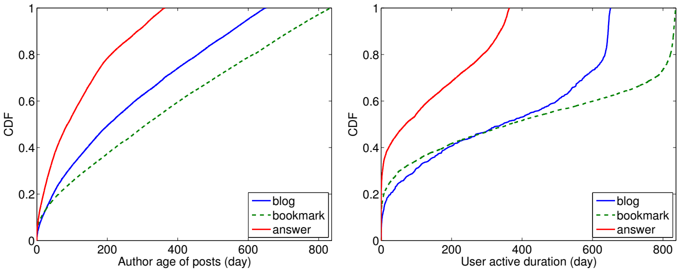
(a) Blog (b) Bookmark

**Figure 2: Daily number of new users over time in OSNs**



(a) Blog (daily) (b) Bookmark (weekly)

**Figure 3: User/post increase rate and overall contribution per user in OSNs**



(a) CDF of author age of posts (b) CDF of user active duration

**Figure 4: User activity over time**

Figure 3(b) shows the *weekly* increase rate of users and posts on the left $y$-axis, and the overall contribution per user on the right $y$-axis for the Bookmark system with time, for the same time period in Figure 2(b). Similar to the user/post increase rate in Blog, the weekly user/post increase rate is defined as the ratio of the number of newly joined users in a week over the cumulative number of users by the last week. Similar to those in the Blog system, both the post increase rate and the user increase rate decrease with time gradually. However, from the figure, we can see that in the week scale, the user increase rate can be greater than the post increase rate, which is very rare in the Blog system. Compared with Figure 2(b), we can see the spikes where the user increase rate is greater than the corresponding post increase rate are caused by the bursts of user joining events, and these joined users have less contributions than common users. Since each of such bursts can last for weeks, the user increase rate actually fluctuates quite big in this time scale, though the general trend is still descending. As a result, as shown in Figure 3(b), the overall contribution per user fluctuates in a time scale larger than weeks, and finally increases linearly at a low rate.

### 3.3 User activity over time and user lifetime

Previous studies assume user's lifetime follows exponential distributions and user activity is uniform over its lifetime or decreases with time exponentially [9, 16]. Understanding user activity over time is important to model the formulation and evolution process of social relationships in a social network, as well as the access and creation traffic of UGC content in a network.

In order to study user posting activity over time, we compute the "author age" of each UGC object in the workloads. The *author age* of a UGC object is defined as the interval from the time when a user joins the network to the time when the object is posted. A uniform distribution of author age of UGC object means user activity is uniform over time, while an exponential distribution of author age of UGC posts means user activity decreases with time exponentially.

Since all three OSNs are expanding over time, the daily number of UGC posting increases with time. To avoid biased estimation, we select users that joined the system in the same week, and extract the author age of each post by these users. Figure 4(a) shows the CDF distribution of the author age of posts in Blog, Bookmark, and Answer OSNs. For the Bookmark OSN, we can see this distribution is almost uniform, meaning most users' bookmarking activities do not change over time due to regular Web browsing. For Blog, posts are a little more concentrated on small author ages than Bookmark, but the main body is close to uniform too since many users tend to post blogs regularly. For Answer, the number of posts with large author ages is even smaller, and becomes uniform when the author age of posting is greater than 200 days. Because answering questions is a kind of altruism behavior, a user may become lazy after providing such services for a certain time duration in the Answer system. Even so, the user activities in all these networks are still not exponentially decreasing with time.

Figure 4(b) shows the CDF of user's active duration, the duration from the user joining time to the last user posting time in our traces. The figure indicates that in all three OSNs, there are a number of users who either have short active durations or have long active durations (especially for Blog and Bookmark). If we assume a user will not return to the OSN after a long inactive time, a short active duration represents a short user lifetime. This means there exist two kinds of users in OSNs: users with short active durations just try the social network system for a short duration and then never or rarely post later; users with long active durations keep posting with time, leading to the uniform body in the CDF of author age of posts.

Our analysis of user activity over time and user lifetime indicates user's lifetime does not follow exponential distribution, while user activity is quite uniform over its lifetime. However, the activity frequency of users may vary significantly, which can be characterized through the distribution of user contributions in OSNs. We present this study in the next section.

## 4. DISTRIBUTION OF USER CONTRIBUTIONS

The workload overview study in the last section provides us some insights on the user activities along time. In this section, we further examine user contributions because in knowledge-sharing oriented social networks, the content contributed by users is the key to attract users and drive the growth of the network.

### 4.1 Original and non-original UGC content

Before we start, we shall clarify that we are concerned about the original content created by users. In general, there are three types of UGC objects in OSNs. The first is the original UGC objects, created by the user who posts them. The second is non-original content obtained through cutting-and-pasting, and the third is advertisement and spam. Since mainly viewers are attracted by the original UGC, we thus focus on the

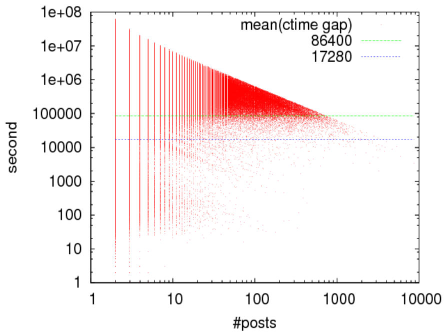


**Figure 5: The average posting intervals of bloggers**

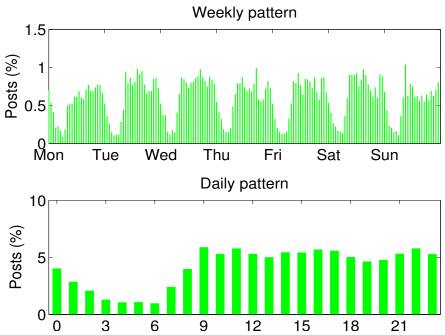


**Figure 6: Weekly/daily patterns of "cut-and-paste" bloggers**

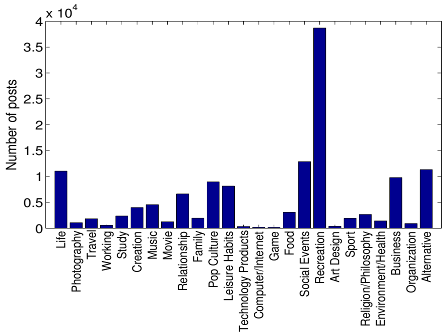


**Figure 7: Forwarded posts in different blog categories**

first type for the study. Therefore, first we need to differentiate and filter out the second and the third types of content from the workloads.

Among the three OSNs we study, Blog mainly contains original UGC and non-original content forwarded from other places. The advertisement and spam contents have been identified and removed using machine learning based approaches when building search index for the system. The advertisement and spam posts in Bookmark and Answer system are filtered out with a similar method as Blog system during search index building period. Furthermore, in Answer, since a user who asks a question can score the answers posted, a person who posts unrelated answers will not earn credits. Thus, we mainly need to filter out non-original content in Blog Article.

Figure 5 shows the average posting interval and the total number of posts for each blogger. We can see there are a small number of bloggers who post a large number of articles with small intervals. These bloggers just cut-and-paste entertainment news and political news from the Web to their own blogspaces. We call these blog posts as forwarded posts. Figure 6 shows the weekly and daily posting patterns of these "cut-and-paste" bloggers. In contrast to Figure 1, the figure shows no concentrated "peak time" for forwarded blog posting, indicating that these users are using "cut-and-paste" at any time in a day. Figure 7 further shows the number of such posts in different article categories (the blog category is selected by the blogger when post). The "recreation" category accounts for most forwarded posts, and the "social events" category ranks the second.

Instead of identifying and removing all forwarded posts, we remove all articles of bloggers whose average posting interval is less than 1/5 day and have posted at least 100 posts altogether, corresponding to a total of 210 users (0.06%) and 1.92% articles. Since our purpose is to study the distribution of the number of original posts by each blogger, which is heavy-tailed, ignoring forwarded posts of users who post a small number of articles (in total) shall not affect the distribution much.

## 4.2 Stretched exponential distribution of user contribution

Since we are interested in the contribution of each user on the original content in an OSN and the variance among different user's contributions, it is natural to rank all users according to their contributions and then identify those with high contributions. If we sort each user by the number of posts in descending order, the function of a user's post number to his/her rank order is called the *rank order distribution function* in the social network. If we normalize the rank order by dividing the total number of users in the social network, then the inverse function of a normalized rank order distribution function is identical to a *complementary cumulative probability distribution function* (CCDF). With the rank distribution, we can focus on those active users who contribute a large amount of high quality UGC content.

The well known Zipf distribution is a rank order distribution, also known as power law. The power law distribution can be expressed as $y_i \propto i^{-\alpha}$ $(1 \leq i \leq n)$, where $y_i$ is the value, $i$ is the rank, and $\alpha$ is a constant. The power law distribution has been widely used in characterizing the Internet, WWW, and social networks.

To analyze the user contribution distribution in depth, Figure 8 shows the distribution of user posts for six types of UGC objects in these three OSNs. In each figure, the $x$ coordinate represents the reference rank of each user, plotted in log scale, while the $y$ coordinate represents the number of UGC objects posted by this user, plotted in both log scale (marked on the right of $y$-axis) and a powered scale (by a constant $c$, as marked on the left of $y$-axis). We call the combination of log scale in $x$ and powered scale in $y$ as the stretched exponential (SE) scale. Note for Blog Article shown in Figure 8(a), the "cut-and-paste" bloggers have been removed. Bookmark Imports in Figure 8(d) represents the bookmarks a user imports from her existing bookmark when joining the Bookmark network. Since in the Answer network, an asker can select an answer for her question as the best answer, we plot the best answers that each user contributes in Figure 8(f), separated from the overall answers in Figure 8(e).

These figures show that in log-log scale, the post rank distributions of users in OSNs have a *flat* head and a *steep* tail, which cannot be fitted with a straight line, indicating they are not power law. However, by selecting a proper constant $c$, all these workloads can be well fitted with a straight line in SE scale. The first several points in Figure 8(c) and 8(d) are much higher than the line predicts, which is called "King effect" [14]. Such a rank distribution is called a *stretched exponential distribution*.

The stretched exponential distribution has been used to characterize the access patterns of Internet media traffic [10]. Its corresponding CCDF function is the Weibull function

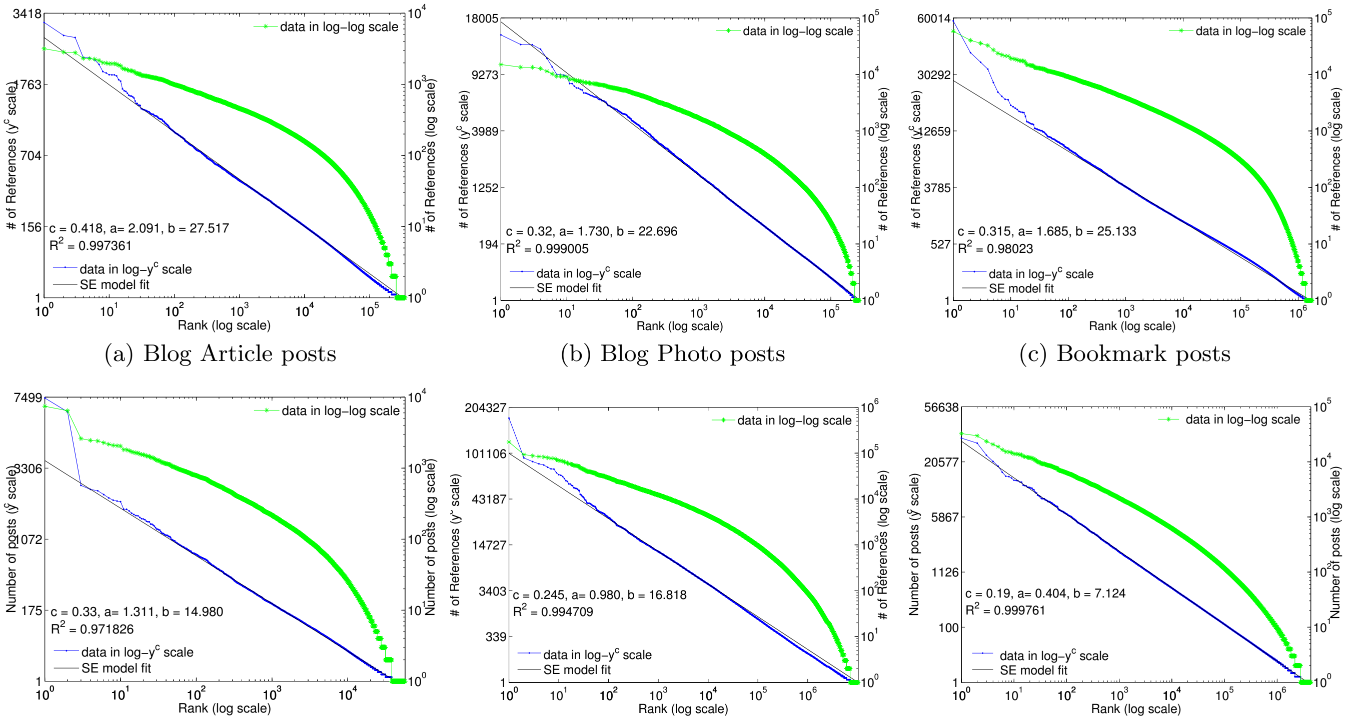
$$P(X \geq x) = e^{-(\frac{x}{x_0})^c}, \qquad (1)$$

where $c$ and $x_0$ are constants. If we rank the $n$ elements in a data set in descending order of the data value $x_i$ $(1 \leq i \leq n)$, we have $P(X \geq x_i) = i/n$. Substitute $x_i$ for $y_i$, the rank distribution function can be expressed as follows

$$y_i^c = -a \log i + b \ \ (1 \leq i \leq n), \qquad (2)$$

where $a = x_0^c$ and $b = y_1^c$. Thus, the data distribution is a straight line in log-$y^c$ plot. If we assume $y_n = 1$, we have

$$b = 1 + a \log n. \qquad (3)$$



(a) Blog Article posts

(b) Blog Photo posts

(c) Bookmark posts

(d) Bookmark imports

(e) Answer posts (overall)

(f) Answer posts (best)

**Figure 8: Stretched exponential distribution of user posts in OSNs**

To get the parameters of a stretched exponential distribution, we use the *maximum likelihood estimation* method (MLE): assuming a data set $\{x_1, x_2, ..., x_n\}$ follows some probability distribution with unknown parameters, the most probable parameters are parameters that make the product of the probability density functions of each element in the data set maximum. Denote the parameter vector as $\theta$, then

$$\theta = \arg\max_\theta \prod_{i=1}^{n} p_\theta(x_i). \tag{4}$$

The probability density function of a Weibull distribution (stretched exponential) is

$$p(x) = c\frac{x^{c-1}}{x_0^c} e^{-(\frac{x}{x_0})^c}. \tag{5}$$

Thus, we have

$$\begin{cases} \frac{1}{c} &= \frac{\sum_{i=1}^{n}(y_i^c \log y_i - y_n^c \log y_n)}{\sum_{i=1}^{n}(y_i^c - y_n^c)} - \frac{1}{n}\sum_{i=1}^{n}\log y_i, \\ a &= \frac{1}{n}\sum_{i=1}^{n}(y_i^c - y_n^c). \end{cases} \tag{6}$$

We first get parameter c with the iteration method, then we get parameter $a$. With Equation 2, parameter $b$ can be estimated as

$$b = \frac{1}{n}\sum_{i=1}^{n}(y_i^c + a\log i). \tag{7}$$

However, in our study, the data to be fit (the number of UGC content a user creates) are positive integers, while the random variables in a Weibull distribution are real numbers. Since there is no data element smaller than one, the parameters given by the MLE method above may result in non-trivial errors in the stretched exponential plot. In order to minimize the model fitting errors caused by the discreteness of data values, especially data elements equal to one, an iterative fitting technique is utilized, described in the following.

We use the *coefficient of determination* of the data fit, also known as $R^2$, as an indicator of fitting errors

$$\begin{cases} SSE &= \sum_{i=1}^{n} w(i)(y_i - (-a\log i + b)^{\frac{1}{c}})^2, \\ SST &= \sum_{i=1}^{n} w(i)(y_i - \overline{y_i})^2, \\ R^2 &= 1 - \frac{SSE}{SST}, \end{cases} \tag{8}$$

where $SSE$ is the sum of weighted squares due to errors, $SST$ is the total sum of weighted squares about the mean, and $w(i)$ is the weight of data point $y_i$. Since the stretched exponential fit is conducted in log scale on the $x$-axis, we select $w(i) = (\log i)' = 1/i$. The closer $R^2$ to 1, the better the model fits the data.

We iteratively truncate the ranked sequence $\{y_i\}$ $(1 \leq i \leq n)$ by removing the last $k$ elements that equals to one in the sequence, then estimate parameters for the truncated sequence using the MLE method, until the $R^2$ value with the estimated parameters is closest to 1 or larger than a threshold. To avoid bias on the fitting error estimation, all elements in the sequence (including the $k$ elements cut off) are considered when computing $R^2$. Our matlab package for stretched exponential fit can be downloaded from `https://sourceforge.net/projects/se-fit-matlab/`.

## 4.3 Model validation

In order to evaluate the goodness-of-fit of the stretched exponential distribution on the data, we conduct Chi-square test as follows. We divide the data value range into $k$ bins $(k \geq 10)$ as evenly as possible, with each bin has at least 5 data points (tail bins are merged when necessary). The Chi-square sum is computed as follows

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}, \tag{9}$$

Table 1: Chi-square test results ($\alpha = 0.05$)

| Data set | $k$ | $\chi^2$ | $\chi^2_{(\alpha,k-c)}$ | Result |
|---|---|---|---|---|
| blog article | 11 | 11.4031 | 14.067 | $\surd$ |
| blog photo | 12 | 14.0723 | 15.507 | $\surd$ |
| bookmark post | 10 | 11.4860 | 12.592 | $\surd$ |
| bookmark import | 11 | 9.3672 | 14.067 | $\surd$ |
| all answer post | 11 | 13.3397 | 14.067 | $\surd$ |
| best answer post | 10 | 7.0005 | 12.592 | $\surd$ |

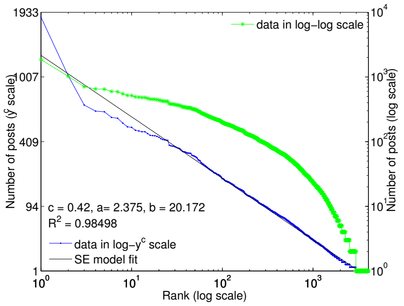


**Figure 9: User posts in Blog (users that join system in the same week)**

where $O_i$ is the observed frequency of $i$-th bin, $E_i$ is the expected frequency of $i$-th bin, and $k$ is the number of bins.

Assume the significance level is $\alpha$ (in our test $\alpha = 0.05$), the assumed distribution is rejected when

$$\chi^2 > \chi^2_{(\alpha,k-c)} \tag{10}$$

where $\chi^2_{(\alpha,k-c)}$ is the Chi-square function, $k$ is number of bins and $c$ is the number of distribution parameters plus 1. The results of our test are presented in Table 1. All 6 data fittings pass the Chi-square test (same as $R^2$ computation, all elements in the data set are considered in Chi-square test). We have also tried power law and lognormal fits on these data sets, however, none of them can pass the Chi-square test.

Figure 8 includes all users in the entire workload. For user contribution in a short duration of each workload, the stretched exponential distribution still holds, with the same stretched factor. However, as we have presented before, the number of users increases with time super-linearly. During the entire workload, some users may become inactive with time. In order to eliminate the effect caused by users of different ages, we select users who join the system in the same week and study their contributions during an entire year. Our results show that the user contributions are still well fit with SE distributions with nearly the same stretch factor $c$ (examined by Chi-square test). Figure 9 shows the fitting results for Blog Article. Due to page limit, we omit other figures.

# 5. IMPLICATIONS OF USER CONTRIBUTION DISTRIBUTIONS

The stretched exponential distribution of user contributions has a number of implications that are different from those based on a power law model. We analyze some of these implications in this section.

## 5.1 The "80-20" rule and "core" users in OSNs

Figure 10 shows the cumulative contribution ratio of top users (over all contribution by all users) in OSNs. As shown in the figure, the cumulative contribution of top users in the tree OSNs roughly follows the so called "80-20" rule: in Blog, the top 20% users account for 80% posts; in Bookmark, the top 16.5% users account for 83.5% posts; in Answer, the top 13% users account for 87% posts.

Table 2: Top-$k$ core users in OSNs

| Data set | $a$ | $c$ | $n$ | $k/n$ | *cumsum* |
|---|---|---|---|---|---|
| Blog Article | 2.091 | 0.418 | 347,710 | 0.148 | 0.7329 |
| Blog Photo | 1.769 | 0.32 | 268,837 | 0.077 | 0.6397 |
| Bookmark | 2.215 | 0.34 | 1,727,353 | 0.083 | 0.6762 |
| Answer | 0.980 | 0.245 | 10,316,931 | 0.047 | 0.6371 |

However, the "80-20" rule of the stretched exponential distribution is different from that of the power law distribution. Consider the cumulative contribution ratio of top-$k$ users in a power law rank distribution $y_i \propto i^{-\alpha}$ and a stretched exponential rank distribution $y_i^c = -a \log i + b$ ($1 \leq i \leq n$ in both distributions), denoted as $T_{se}$ and $T_{pow}$, respectively. Figure 11 shows the comparison of $T_{se}$ and $T_{pow}$ with log scale in $x$ axis. The parameters of the SE plot are based on the blog article data, while the skewness factor of the power law plot, $\alpha$, is set to 0.9. We can see although the two curves intersect at the "80-20" point, for a small $k$, the cumulative contribution ratio of top $k$ users in a stretched exponential network is much smaller than that in a power law network.

When $n \to \infty$, for a limited value of $k$, we can prove

$$\frac{T_{se}}{T_{pow}} = \lim_{n\to\infty} \frac{k}{\sum_{i=1}^{k} \frac{1}{i^\alpha}} \frac{(1 + a \log n)^{\frac{1}{c}}}{a^{\frac{1}{c}} \Gamma(1 + \frac{1}{c})(1 - \alpha) n^\alpha} = 0. \tag{11}$$

The analysis above indicates that in contrast to a power law distribution, a stretched exponential distribution is less skewed, meaning a small number of top users cannot dominate the network as those in power law networks. This can be reflected from the log-log plot of the user contribution rank distribution. As shown in the log-log plot of Figure 8, the user contribution distribution curve has two modes in general. The first mode is quite flat, corresponding to a small number of top users, where the change rate of user contribution decreases with the change rate of user rank $k$ slowly. The second mode is much steeper, corresponding to the majority of users, where the change rate of contribution decreases with the change rate of rank $k$ significantly.

This observation motivates us to identify those "core" users and the corresponding contributions in a social network with the top-$k$ analysis. For this purpose, we select a rank $k$ so that for users with a rank $i \leq k$, the decrease rate of user contribution is smaller than the increase rate of user rank. Thus for $i = k$, we have

$$\frac{dy(i)}{y(i)} + \frac{di}{i} = 0, \text{ i.e., } \frac{d \log y(i)}{d \log i} = -1. \tag{12}$$

Let $X = \log i$ and $Y = \log y$. Figure 12 shows a stretched exponential distribution curve in log-log scale. Line $AB$ with slope $-1$ is tangent to the SE curve at point $(X_0, Y_0)$. So the geometric meaning of $k$ is that $k = \exp(X_0)$, and we have

$$X_0 = \frac{b}{a} - \frac{1}{c}. \tag{13}$$

Assuming $b = 1 + a \log n$ (Equation 3), we have

$$\frac{k}{n} = \exp(\frac{1}{a} - \frac{1}{c}). \tag{14}$$

Table 2 shows the number of top-$k$ core users for different kinds of UGC objects in OSNs. Column *cumsum* is the cumulative contribution ratio of total content in the OSN for top-$k$ core users. In general about 5% to 15% users can be considered as core users in OSNs. However, the *cumsum* of

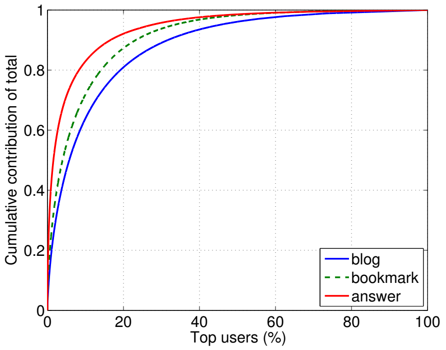


**Figure 10: Cumulative contribution ratio of top users in OSNs**

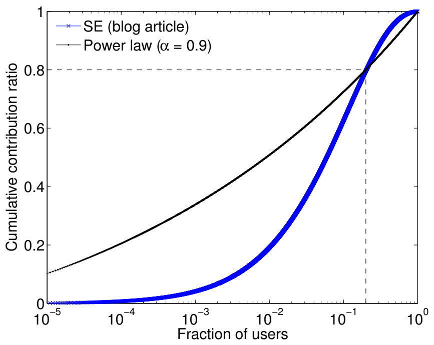


**Figure 11: Contribution of top users in SE and power law networks**

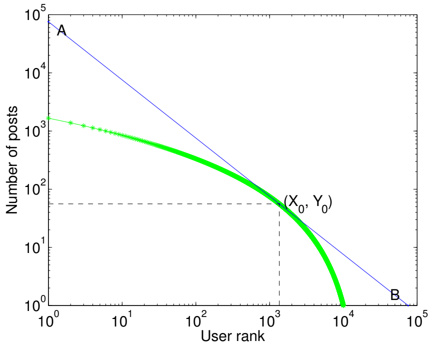


**Figure 12: Top user selection**

these top-$k$ core users are quite close for all cases, ranging from 63% to 73%. The same method can be applied to estimate "popular" objects in content sharing OSNs where the request patterns follow stretched exponential distributions.

## 5.2 Different UGC creation patterns in OSNs

The "80-20" rule of user contribution in OSNs indicates a small fraction of users contribute most content in the network. However, this metric is quite rough, and cannot reflect the inequality of users' knowledge contributions in OSNs. The top-$k$ analysis based on stretched exponential model provides a quantitative method to characterize the concentration of UGC contributions from different users.

To further understand the inequality of users' contributions on UGC objects of different types, we conduct the stretched exponential fit on different durations of the three data sets. Our results show that, although different durations of data set have different numbers of users and UGC objects, parameter $c$ is almost a constant for the same social network system and content type across different durations, while parameter $a$ varies for different durations.

We have also studied different classes of objects in the Blog social network, by considering blog articles of different sizes and different numbers of tags attached by the authors. The stretched exponential parameters of different UGC objects are listed in Table 3. As shown in the Table, the parameter $c$ for best answer posts is smaller than that of all answer posts, and the parameter $c$ for blog photos is smaller than that of blog articles. In contrast, for Bookmark, the parameters $c$ for bookmark posts (bookmarks generated with Bookmark plug-in) and imported bookmarks (bookmarks generated with Web browser, which are imported to the Bookmark when a user joins the system) are almost the same. We conjecture that parameter $c$ reflects the *quality* of a UGC object or the *effort* of creating a UGC object, which may characterize some intrinsic property of UGC objects in social networks: the more effort a user needs to make to create a UGC object, the smaller $c$ is. In the Answer system, the "best answer"s are selected by the user who asked the question, thus the content quality has been judged by the asker herself. If we assume high quality answers needs more effort to create than low quality answers, then the distribution of user contribution for best answer posts, which are effort-consuming, would have smaller $c$ than that for normal answer posts. For Blog system, it is easy to understand that longer articles need more effort to compose than shorter articles, and selecting tags for an article needs extra effort. It is also understandable that composing a short blog article in the Web browser needs less effort than taking a photo, transferring it to a computer, making some edit work such as rescaling, and then uploading it to one's blogspace with some text descriptions. On the other hand, it is straightforward that there is no significant difference between the effort of bookmarking a Web page using a Web browser and that using a Bookmark plug-in. However, it is hard to compare Blog/Answer posts and Bookmark posts directly since the posting mechanisms and user contributions of these two kinds of UGC objects are significantly different.

**Table 3: Stretched exponential parameters for different types of UGC contents**

| UGC workload | $c$ | $a$ |
|---|---|---|
| Blog Article (all posts) | 0.418 | 2.091 |
| Blog Article (with tags) | 0.3 | 0.761 |
| Blog Article (> 1 KB) | 0.39 | 1.451 |
| Blog Article (> 2 KB) | 0.31 | 0.702 |
| Blog photo | 0.32 | 1.730 |
| Bookmark (imports) | 0.33 | 1.311 |
| Bookmark (all posts) | 0.315 | 1.685 |
| Answer (all posts) | 0.245 | 0.980 |
| Answer (best answers) | 0.19 | 0.404 |

For Wikipedia, the effort of composing a Wikipedia article can be much greater than that of composing a good answer for a user asked question. If we assume that the user contribution in Wikipedia also follows the stretched exponential distribution, then the parameter $c$ would be much smaller than that of best answer in our study (0.19). When $c$ is small, we have $y^c \sim \log y$, and Equation 2 becomes power law. Thus, the reported power law distribution of Wikipedia author contribution [23] can be explained.

For the emerging microblog OSNs, such as Twitter (`http://twitter.com/`), the user participations are pervasive as the effort for participating is decreasing in general. Based on our study, we conjecture that the SE distribution in such OSNs should become flatter as the contributions of normal users become dominant in the network.

Our study suggests that the SE distribution can more accurately reflect individual user contributions in these OSNs (and potentially all knowledge-sharing OSNs), which is significantly different from the power law distributions that hold for other properties of social networks, such as user's online connections in IM and email networks and in-bound degree of blog networks. These phenomena are caused by the aggregation effect of multiple users, which can be explained by the "rich-get-richer" or preferential attachment model [16]. However, for the activity of individual users, the lack of aggregated "rich-get-richer" effect implies power law cannot hold, including its variants such as the power law with exponential cutoff model.

The distribution of individual user contribution is the building block to model more complex social network phenomena. Although models have been proposed to describe how user links are created in social networks and how the user networks evolve with time, the process of user link initialization during the evolution is often oversimplified, as we mentioned in Section 3.3. The stretched exponential model can provide in-depth understanding on these social network phenomena.

### 5.3 Discussion: UGC production vs. UGC consumption

Although we have not analyzed the click rates of UGC objects in OSNs in this work, a study of Internet media access patterns has shown that the reference rank distribution of objects in Internet media systems, including viral video social networks such as YouTube, follows the stretched exponential distribution [10]. In contrast to UGC creation in social networks, which is to *produce* content, a user request in video social networks is to *consume* content. For user requests in a video system, the stretch factor represents the median file size of an object or the average length of an object that users view, i.e., the amount of gain one obtains. The larger the file size (measured by the length viewed) is, the greater the stretch factor would be. For user contributions in a social network, the stretch factor represents the quality of a UGC object or the effort to create an object. The more efforts to create an object we make for the higher quality of the created object, the smaller the stretch factor would be. What is the relationship between UGC creation and UGC consumption in a social network? Why some social networks are successful, with high user population and high quality content, some are not? Can we predict the page views or traffic volume of a social network, with a few properties of UGC creation patterns and consumption patterns in the network? A deeper understanding of user behavior patterns in OSNs can help us understand the driving force in these networks, design effective participation mechanisms for social applications, and provide efficient resource management for underlying supporting systems.

## 6. CONCLUSION

Technology advancements have brought up many OSNs on the Internet. For knowledge-sharing oriented OSNs, the user activities and contributions are critical. In this work, we have extensively analyzed user activities and contributions in three large OSNs and have revealed several new findings that are different from or contradicting to common assumptions. In particular, the user lifetime in these OSNs does not follow exponential distributions, and the user contribution does not follow power law distributions, but stretched exponential. Furthermore, different types of UGC content have different characteristics under the stretched exponential model. Our results provide timely insights for the current social network industry and research communities, and lay out a solid foundation to guide the design, modeling, and simulation of OSNs with different properties and scales.

## Acknowledgments

This work has been supported in part by U.S. NSF under grants CNS-0509061, CNS-0621629, CNS-0621631, and CNS-0746649, and by U.S. AFOSR under grant FA9550-09-1-0071. We appreciate encouragement and constructive comments from Nam Nguyen and anonymous referees.

## 7. REFERENCES

[1] L. Adamic and N. Glance. The political blogsphere and the 2004 U.S. election: Divided they blog. In *Proc. of Workshop on Link Discovery*, 2005.

[2] E. Adar and L. Adamic. Tracking information epidemics in blogspace. In *Proc. of IEEE/WIC/ACM International Conference on Web Intelligence*, 2005.

[3] N. Bailey. *The Mathematical Theory of Infectious Diseases and its Applications*. Griffin, London, 1975.

[4] F. Bell. The rise of user-generated content. http://www.entrepreneur.com/technology/managingtechnology/web20columnistfrankbell/article183432.html, 2007.

[5] X. Cheng, C. Dale, and J. Liu. Statistics and social networking of YouTube videos. In *Proc. of IEEE IWQoS*, 2008.

[6] M. Crovella and A. Bestavros. Self-similarity in world wide web traffic, evidence and possible causes. In *Proc. of ACM SIGMETRICS*, 1996.

[7] M. Gjoka, M. Sirivianos, A. Markopoulou, and X. Yang. Poking Facebook: Characterization of OSN applications. In *Proc. of ACM SIGCOMM WOSN*, 2008.

[8] D. Gruhl, R. Guha, D. Liben-Noewll, and A. Tomkins. Information diffusion through blogspace. In *Proc. of WWW*, 2004.

[9] L. Guo, S. Chen, Z. Xiao, E. Tan, X. Ding, and X. Zhang. Measurements, analysis, and modeling of BitTorrent-like systems. In *Proc. of ACM SIGCOMM IMC*, 2005.

[10] L. Guo, E. Tan, S. Chen, Z. Xiao, and X. Zhang. The stretched exponential distribution of Internet media access patterns. In *Proc. of ACM PODC*, 2008.

[11] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *Proc. of ACM SIGKDD*, 2003.

[12] A. Kittur, E. Chi, B. Pendleton, B. Suh, and T. Mytkowicz. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. In *Proc. of ACM CHI*, 2007.

[13] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *Proc. of WWW*, 2003.

[14] J. Laherrere and D. Sornette. Stretched exponential distributions in nature and economy: "fat tails" with characteristic scales. *European Physical Journal B*, 2:525–539, 1998.

[15] J. Leskovec, L. Adamic, and B. Huberman. The dynamics of viral marketing. In *Proc. of ACM Electronic Commerce*, 2006.

[16] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *Proc. of ACM SIGKDD*, 2008.

[17] J. Leskovec and E. Horvitz. Planetary-scale views on a large instant-messaging network. In *Proc. of WWW*, 2008.

[18] J. Leskovec, K. Lang, A. Dasgupta, and M. Mahoney. Statistical properties of community structure in large social and information networks. In *Proc. of WWW*, 2008.

[19] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs. In *Proc. of SIAM Conference on Data Mining*, 2007.

[20] M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. In *Proc. of ACM SIGKDD*, 2002.

[21] M. Seshadri, S. Machiraju, A. Sridharan, J. Bolot, C. Faloutsos, and J. Leskovec. Mobile call graphs: Beyond power-law and lognormal distributions. In *Proc. of ACM SIGKDD*, 2008.

[22] D. Stutzbach and R. Rejaie. Understanding churn in peer-to-peer networks. In *Proc. of ACM SIGCOMM IMC*, 2006.

[23] J. Voss. Measuring wikipedia. In *Proc. of ISSI*, 2005.

[24] J. Wales. Wikipedia, emergence, and the wisdom of crowds. http://mail.wikipedia.org/pipermail/wikipedia-l/2005-May/039397.html, 2005.

[25] M. Wang, T. Madhyastha, N. Chang, S. Papadimitriou, and C. Faloutsos. Data mining meets performane evaluation: Fast algorithms for modeling bursty traffic. In *Proc. of ICDE*, 2002.

[26] D. Watts. A simple model of global cascades on random networks. *PNAS*, 99:5766–5771, 2002.

[27] Z. Xiao, L. Guo, and J. Tracey. Understanding instant messaging traffic characteristics. In *Proc. of IEEE ICDCS*, 2007.