From User Comments to On-line Conversations

Chunyan Wang
Department of Applied Physics
Stanford University
California, USA
chunyan@stanford.edu

Mao Ye
Social Computing Group
HP Labs
California, USA
mxy177@cse.psu.edu

Bernardo A. Huberman
Social Computing Group
HP Labs
California, USA
bernardo.huberman@hp.com

ABSTRACT

We present an analysis of user conversations in on-line social media and their evolution over time. We propose a dynamic model that predicts the growth dynamics and structural properties of conversation threads. The model reconciles the differing observations that have been reported in existing studies. By separating artificial factors from user behavior, we show that there are actually underlying rules in common for on-line conversations in different social media websites. Results of our model are supported by empirical measurements throughout a number of different social media websites.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and Behavior Sciences; G.3 [**Mathematics of Computing**]: Probability and Statistics

General Terms

Human Factors, Theory, Measurement

Keywords

conversation dynamics, social networks

1. INTRODUCTION

The rapid development of social media websites has dramatically changed the way that people communicate with each other. A particularly interesting phenomenon is the prominent role of users as a leading information source within these websites. For example, various on-line media and review sites provide commenting facilities for users to exchange opinions and express sentiments about news, stories and products. These user-generated comments link together and form a conversation thread, which is essentially a distinctive kind of information network that has a life span significantly shorter than other information networks.

KDD'12, August 12–16, 2012, Beijing, China.

As pointed out in [1], despite the significant research efforts made on information networks, dynamics of conversation threads have not received enough attention so far. As a matter of fact, the dynamics of such linked information plays a fundamental role in opinion spread and formation [2, 3], word-of-mouth effects [4] and collective problem solving [5,6]. Existing empirical studies on on-line conversations seem to yield conflicting results about the basic statistical properties: while some results demonstrate that the size distribution of posts and reviews follow a heavy-tailed distribution such as Zipf's law [1,7] or lognormal distribution [8,9], another portion of the literature suggest a light-tailed one, such as negative binomial [10, 11]. A fundamental question is how can two apparently different categories of distributions describe the same type of information network? And what are the dominating factors that are responsible for the observed differences? In this paper, we focus on addressing these problems by proposing a dynamic model for on-line conversations. The contributions of our study can be summarized in three complementary dimensions:

- **User Attention on New Items.** We examine the dynamics of user attention on new items. We analyze the duration of new topics displayed to users, and also the non-Poisson nature of user commenting behavior.
- **Model of On-line Conversations.** We propose a dynamic model for conversation growth based on a number of different factors, including the exposure duration of topics on the website, patterns of users' commenting behavior, and also the impacts of social propagation and social influence. The model successfully reconciles existing discrepancies in reported studies, and also explains the structural properties within a conversation thread.
- **Size and Structure of Conversations.** We compare results from our model with empirical measurements using datasets from Digg[1], Reddit[2] and Epinions[3].

The rest of this paper is organized as follows. Section 2 reviews the related work. Section 3 introduces the datasets used in our empirical studies. Section 4 introduces key observations on user behavior and the impact of featuring mechanisms on user attention. Based on these observations, Section 5 introduces the dynamic model for user conversation.

[1]http://digg.com
[2]http://reddit.com
[3]http://epinions.com

Section 6 compares predictions of the model with empirical measurements. Section 7 concludes the paper with a discussion.

2. RELATED WORK

In this section, we discuss some of the relevant studies categorized in three sub-areas.

2.1 Information Spread and Conversations

In the field of information dissemination, the pioneering work of Liben-Nowell and Kleinberg [12] modeled information spread as a propagation of chain letter. Golub and Jackson [13] extended this work with a branching process model combined with the selection bias of observation. In social media, Leskovec et al. [14] investigated the propagation of memes across the Web. The main concern of these studies is to understand information spread in the context of social network. Other relevant studies focused on properties of information networks that are formed in the process of information spread. Mishne et al. [7] investigated weblog comments for identifying blog post controversy, Duarte et al. [15] engaged in describing blogsphere access patterns from the blog server point, Kaltenbrunner et al. [8] measured community response time in terms of comment activity on Slashdot stories, Choudhury et al. [16] characterize conversations through their interestingness, and finally, Kumar et al. [1] modeled the dynamics of conversations with a branching process incorporating recency. Along with the increasing results in on-line conversations, one emerging problem is the seemingly conflicting observations even about the most basic statistical properties of conversation threads. As mentioned earlier, some studies suggest that the size follows a heavy-tailed distribution[4] such as Zipf's law [1, 7] or lognormal distribution [8, 9]. Other measurements point to a Poisson family [10, 11]. One main focus of this paper is to provide an explanation for these differing observations.

2.2 Dynamics of User Attention

Another set of studies related to our work is the dynamics of user attention. In the information age, posts on websites compete with each other for the scarce attention of users [17, 18]. To help users find high quality content, social media websites usually place information in a "featured column" for popular items, such as the "popular" page on Youtube, the "trending" column on Twitter, the "what's hot" column on Reddit. Existing studies show that this featuring mechanism has significant impact on user attention [19]. To enter the featured column, posts need to reach a threshold of critical mass. Studies on Digg [20, 21], Youtube [22], Wikipedia [23] and Twitter [4, 24] successfully explain the attention dynamics of topics after the critical mass threshold. However, it is still unclear about the attention dynamics of the vast majority of topics and stories before reaching the critical mass. As such, it has remained an open question about the attention dynamics and the initial growth of these items. We attempt to propose dynamics of the user attention, measured in the number of user comments, for these general items on social media websites.

[4]In this paper, we use the term heavy-tailed to denote the probability distributions whose tails are not exponentially bounded, i.e. $\lim_{x\to\infty} e^{\lambda x} P(X > x) = \infty, \lambda > 0$.

2.3 Dynamics of Human Behavior

Traditionally, the dynamics of human behavior is described by a series of Poisson process events under the context of Internet [31]. Recent advances in human behavior suggest that the waiting time between two consecutive events follows power law scaling, ranging from email exchanges [25–27] to web browsing [22, 28–30]. Various new models have been proposed to interpret the observed scaling of waiting times [25, 30, 32]. While most existing studies emphasize on explaining the nature and origin of the observation, the implications of this power law scaling on information spread and attention dynamics has not been exploited thus far. In this paper, we examine the waiting time of human comments and more importantly, we use the power law scaling of human behavior to understand the dynamics of on-line conversation.

3. DATA

Three datasets from Digg, Reddit and Epinions are used in our empirical measurements. To collect these datasets, we monitored the website for newly created items or topics. We kept track of these topics' user comments for a time span of at least three months since the topics' creation, to make sure that the growth saturates. We also recorded related information such as the time stamp when the topic is removed from the column for displaying new items. In our empirical studies, we perform the same treatments on these datasets whenever possible.

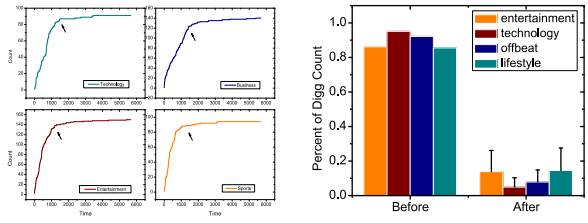
Digg is an interactive social media website, which allows its users to share and comment on news and stories. Users of the website select and direct attention to a few items from a very large pool of submissions. They can read, Digg, Bury, and leave comments on the topic or other users' comments. In our study, we monitored the website for a total number of 17,322 topics containing 158,782 comments. Each comment was labeled by its posting time. To obtain information about individual user's commenting behavior, we also monitored a number of 8,616 users on Digg and collected all these users' comments.

Another dataset used in our study was from the social news website Reddit. Users on Reddit submit content in the form of either a link or a text post. Other users reading the post can express their opinions by commenting on the original post. Similar to Digg, comments on Reddit can also be directed to existing comments. In our study, we collected over 78,312 comments from 8,428 conversation threads. For each comment, we recorded the user-id and timestamps of the comments. We also recorded to which comment or post that comment is referring.

To ensure that our observations are not limited to news media sites, we included a dataset of consumer review from Epinions. Epinions is a who-trust-whom consumer review site, and users write their personal reviews on a wide variety of products, ranging from automobiles to media (music, books, movies and etc.). Members of the site can decide whether to trust other members based on their reviews. Again, every user on the website can comment on the reviews or on the existing comments. We collected 88,859 unique users' comments from the website. We also collected 286,317 topics from different categories containing a total of 722,475 user comments.

4. USER ATTENTION TO NEW ITEMS

To understand the underlying mechanisms governing user attention and on-line conversations, we first look at the growth of attention on newly generated items in social media. Figure 1 (a) shows the growth of cumulative user attention measured in Digg count of four typical topics from Digg. Results from Reddit and Epinions are similar to the one in Digg. One general observation for topics from different categories and different websites is that the cumulative count saturates to a point where a sharp drop of the growth rate is apparent. We explain this observation with the following reasoning. To help users explore new topics, typical social media websites place newly generated topics in the "upcoming" and "new" columns since their creation time. Users visit the website regularly and discover these newly generated stories. After a period of time, the old topics are replaced with newly generated contents. While these replaced items can still be accessed through search queries, it has significantly less chance to be exposed to general users. So it explains why the growth of attention eventually saturates. To confirm this explanation, we kept track of the time when the topics are removed from the front page of "upcoming" column on Digg. We find that the saturation point has a high correlation with the time point when the topic is removed. The black arrow in Figure 1 (a) identifies the time point of removal, which is very close to the saturation point. Figure 1 (b) compares the number of user comments happened before and after the inflection point. The averaged percentages of comments happened before the inflection points are 0.8616, 0.9509, 0.9215 and 0.8548 respectively for categories of entertainment, technology, offbeat, and lifestyle. Different colors in the plot represent different sub-categories. Error bars in the plot indicate one standard deviation of the data in the sub-category. As expected, most of the comments are generated before removing from the "upcoming" column. In



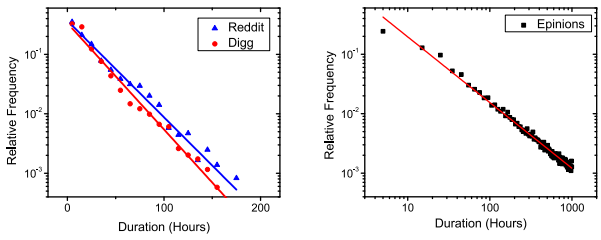
(a) Cumulative Size (b) Ratio

Figure 1: (a) User attention as a function of time (minutes) for four typical topics on Digg. The black arrow in each plot shows the inflection point where the topic is removed from the "upcoming" column. (b) Percentage of user comments happened before and after the inflection point.

the rest of this paper, we name the point, when a topic is replaced from "new" or "upcoming" column, as "inflection point". We name the duration that a topic stays in the column as "exposure duration". The exposure duration varies from topic to topic, which is largely determined by the speed of generating new items and the hidden algorithms used by the website to remove old ones. There are two important factors that are dominating the conversation growth before saturation: (i) the length of exposure duration and (ii) the patterns of user commenting behavior. Next, we investigate these two factors in detail separately.

4.1 Distribution of Exposure Durations

The duration of items placed in the "new" column since creation plays a fundamental role in the initial growth of attention dynamics and comment counts. Here, we empirically measure the distribution of this exposure duration from three mainstream social media websites Digg, Reddit and Epinions.



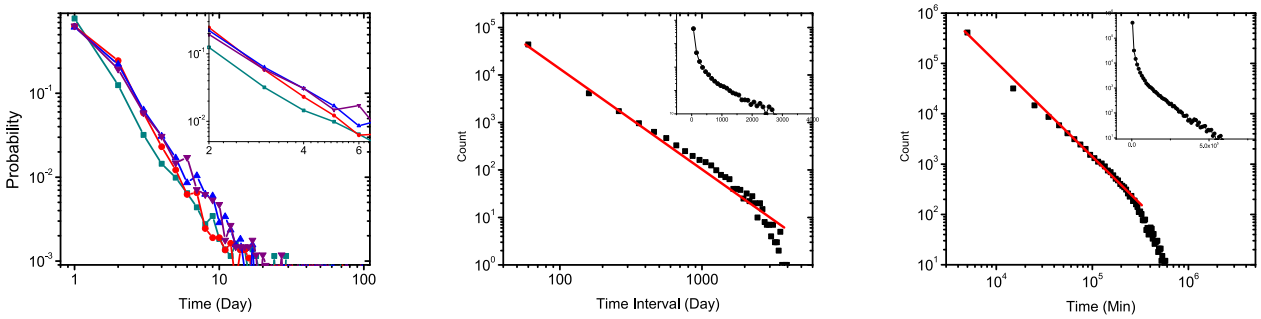
(a) Digg & Reddit (b) Epinions

Figure 2: Density plot of exposure duration for new topics. (a) Digg and Reddit, (b) Epinions.

On Digg, there is a specific column named "upcoming news" for newly generated items. Topics in this column are sorted by creation time, with the newest item ranking on the top of the page. When new topic comes, all of the existing items move downwards on the web page. In doing so, topics of the website would fade away from users' attention gradually. Here we measure the duration that the item maintains on the first top 50 items in the "upcoming news' column. Similar results are observed when we change this threshold limit. On Reddit and Epinions, we use similar gathering methodologies and treatments. In Figure 2 (a), an exponential distribution can be observed from the semi-log plot for both of Digg and Reddit. And for Epinions in Figure 2 (b), a Pareto distribution for the exposure duration is observed from the straight line in log-log plot. Since the exposure duration is determined by various specific factors such as the speed of item generation and the hidden algorithms used, different distribution of exposure durations are expected for websites with similar appearances. For instance, if items in the column are removed with a fixed probability at each time step, the duration is expected to be geometrically or exponentially distributed [24]. If items are pushed down a fixed-size list until falling off, the duration would follow Erlang distribution, which is the sum of exponentials. Various other optimizing strategies can result in a power law distribution or a lognormal distributed durations [33]. For this reasoning, one could not presume the distribution of exposure duration without knowledge about the hidden algorithms or empirical measurements. The impact of this observed differences is later discussed in the model section.

4.2 Patterns of User Commenting Behavior

In last sub-section, we focus on the side of websites, looking at the distribution of for how long a new item is exposed to general users. Now, we turn our attention to users'



(a) Individuals (b) Epinions (c) Digg

Figure 3: Density plot of waiting times between two consecutive comments from a user, (a) time intervals of four users with largest comment count on Digg, (b) intervals from all users in Epinions dataset and (c) intervals from all users in Digg dataset. The upper-right plot in (a) shows a zoomed-in view of density plot. The red straight line in (b) and (c) suggests a power-law family of distribution. The plot on the upper right corner in (b) and (c) demonstrates data in a semi-log scale.

commenting behavior, i.e. the distribution of waiting times between two comments from the same user.

First, we look at the distribution of two consecutive comments from single users. Figure 3 (a) demonstrates the distribution of waiting times for four typical users on Digg in a log-log scale. The upper right plot in the figure shows the scaling region ranging from 2 to 6 days. One interesting observation from the plot is that the four colored lines, despite coming from different users, show similar scaling relationship. And the slope of the line also varies little in the four samples with the largest comment count in our dataset. Similar scaling is observed for other users of the website. This observation suggests that different users share similar patterns of commenting behaviors. So we turn our attention to study the behavior of aggregated users on a whole, by treating users as identical. We empirically measure the distribution of waiting times by collecting the time series data of all comments from users. The density plot of waiting times between two consecutive comments in a log-log scale is shown in Figure 3 (b) and (c). The red straight line in the plot suggests a power-law scaling of waiting times distribution. The plots on the upper right corner demonstrate the same data but in a semi-log scale. From these two plots, the distribution clearly deviates from an exponential distribution. The cutoff for (b) and (c) can be explained by the finite-size effect. The above observations suggest that the commenting behavior of human can not be described by a Poisson process as assumed in prior studies [31]. We find that the density plot is best fitted with an upper-truncated Pareto distribution. Based on the maximum-likelihood estimation (MLE) approach [34] for upper-truncated Pareto distribution, the exponent for Epinions is estimated to be -1.5670, when the lower bound is set to equal one unit and the upper bound is set to be equal to the largest observation in our records. Similarly for the MLE of Digg dataset, the exponent is estimated to be -1.1262. This result implies that, for each user, frequent comments may follow by a significantly long period of inactivity. In the following, we explore the implications of this non-Poisson nature of human behavior.

5. MODEL OF ON-LINE CONVERSATIONS

We introduced basic properties about the duration of new topics getting displayed to users and the patterns of user behavior. In this section, based on these properties, we propose a model for the growth dynamics of on-line conversations. We show that the basic scaling relationship deduced from our model is a robust one, in that the scenario discussed can arise under very simple assumptions.

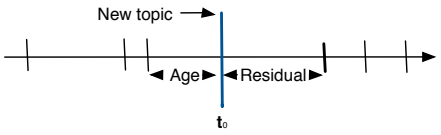
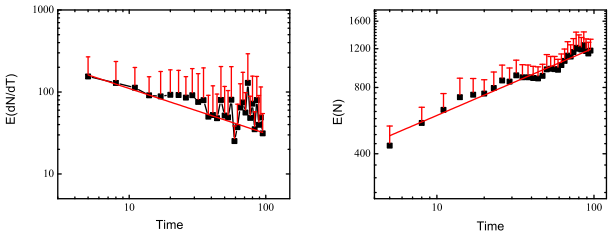
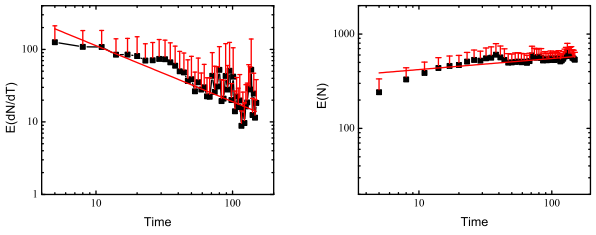


Figure 4: The arrival pattern of comments from one user. Every short vertical line in the figure represents the time of a comment from the user. The blue thick vertical bar represents the time point when a new topic is released.

Based on results form Section 4, we assume that users check the websites' "new" columns regularly to discover topics. In our model, we use t_0 to denote the time point when the topic is created. We use t to denote the time passed since the creation of the topic, and T to denote the exposure duration for a topic. We use $N(t)$ to denote the cumulative count of user comments on a topic, or the size of conversation. The waiting time of two consecutive comments from a user follows a upper truncated Pareto distribution. Here, we simplify the problem by assuming that users share the same microscopic behaviors, i.e. the waiting time for different users comes from the same distribution. In doing so, we are able to model the process of M users as M independent concurrent counting process, so it is sufficient to consider the case of one individual user. For that user, the waiting time between two comments is an independent and identical variable. The counting process of that specific user forms a renewal process, as depicted in Figure 4. Here, we let x_i denote the inter-arrival time of the ith comment from the user, $Y(t_0)$ denote the time from t_0 until the next renewal,



(a) Digg dN-dT (b) Digg N-dT



(c) Reddit dN-dT (d) Reddit N-dT

Figure 5: Comment growth dynamics on Digg in log-log scale. (a) dN/dT as a function of T, the red line in the figure shows the linear fit of data in log-log scale, with a slope of -0.5483 and a standard error of 0.3809. (b) N as a function of T, the red line in the figure shows the linear fit of data in log-log scale, with a slope of 0.3066 and a standard error of 0.046. Comment growth dynamics on Reddit in log-log scale. (c) dN/dT as a function of T, the red line in the figure shows the linear fit of data in log-log scale, with a slope of -0.7720 and a standard error of 0.2454. (d) N as a function of T, the red line in the figure shows the linear fit of data in log-log scale, with a slope of 0.1172 and a standard error of 0.0057.

$A(t_0)$ denote the time from t_0 since the last renewal. If the waiting time of each user's comments follows an independent and identical upper-truncated Pareto distribution, the distribution of the interval can be written as

$$f(x) = \frac{c}{a^{-c} - b^{-c}} x^{-c-1}. \tag{1}$$

Here, a is the lower bound of the time interval required for a user to post another comment. b is the upper bound of the Pareto distribution stemming from its finite size effect. The cumulative distribution of the truncated power-law takes form of $F(x) = \frac{a^{-c} - x^{-c}}{a^{-c} - b^{-c}}$, for $a \le x \le b$. We have

$$\overline{F(x)} = 1 - F(x) = \begin{cases} 1, & x < a \\ \frac{x^{-c} - b^{-c}}{a^{-c} - b^{-c}}, & a \le x \le b \\ 0, & x \ge b \end{cases} \tag{2}$$

For an individual user, the quantity of interest is the probability that a comment happens at $Y(t_0)$ on the specified topic, as $t_0 \to \infty$. To derive this, we begin with the probability of the user commenting on any of the existing topics in the column at time $Y(t_0)$. Since the inter-arrival time is independent and identically distributed, $Y(t_0)$ and $A(t_0)$ form an alternating renewal process. From results of the key renewal theorem [35], we have

$$\lim_{t_0 \to \infty} P\{Y(t_0) \le y\} = \frac{1}{\mu} \int_0^y \overline{F(x)} dx, \qquad (0 \le y \le b). \tag{3}$$

In this equation, $\mu = \frac{c}{c-1} \frac{a^{-c+1} - b^{-c+1}}{a^{-c} - b^{-c}}$ is the expectation of random variable x. $A(t_0)$ and $Y(t_0)$ do not have to share the same distribution, due to the impact of the lower bound a. By taking Equation 2 back to Equation 3, we obtain the probability of user comment on any of the items at $Y(t_0)$,

$$P\{Y(t_0) = y\} = \frac{\partial P(Y(t_0) \le y)}{\partial y} \sim y^{-c}. \tag{4}$$

Based on our assumptions, users can choose to comment on one topic from the column at a time. So at t_0, the user may choose to comment on any of the existing topics that are still in the column. Neglecting all other factors, if we assign a fixed probability α for the user to choose the specified topic from all topics in the column, the size of conversation scales with $N(t) \sim \alpha t^{-c+1}$. One insight of this equation is that the probability of one more additional comment adding to the topic inversely scales with time. The interestingness measurement α can be assigned different values to different topics. To derive the most common properties of conversation growth dynamics, we fix α over topics in our model.

Thus far, we have derived the growth of conversation size without considering the resonating nature of on-line conversations. Now we take these important characteristics into consideration by writing α as $\alpha(N(t))$. The reason that α is a function of $N(t)$ comes from existing works in information cascades and social influence. The intuitive understanding is that the more popular a topic is, the more likely that a user comment on it or come back to comment again. Given that N scales with t, we assume $\alpha = \gamma t^{c_0}$, γ is a constant factor. c_0 is a positive exponent measuring the combined impacts of factors such as resonance and social influence. In the extreme of $c_0 = 0$, α would be a constant, when there is no other impacts such as social influence. Now we combine existing two parts together to derive the dynamics of conversation size growth. Noting that the expected number of increment of comments at time point t would be proportional to $\gamma t^{c_0} M t^{-c}$. The total number of comments for a given topic, N, grows like

$$\frac{dN(t)}{dt} = \gamma M t^{-c+c_0}. \tag{5}$$

Thus, $N(t) \sim t^{-c+c_0+1}$, i.e. $ln(N(t))$ scales linearly with $ln(t)$. To confirm this derivation, we compare this result with the empirically measured growth of conversation size. As shown in Figure 5, the plot in log-log scale shows the expected scaling relationship between time and number of increments. In the figures, the red line in the figure shows the linear fit of data in log-log scale. For Digg, linear fitting gives a slope of -0.5483 and a standard error of 0.3809 as shown in Figure 5 (a) and a slope of 0.3066 with a standard error of 0.046 for Figure 5 (b). For Reddit, linear fitting yields a slope of -0.7720 and a standard error of 0.2454 for Figure 5 (c) and a slope of 0.1172 with a standard error of 0.0057 for Figure 5 (d). The two estimated exponent values obtained from two fittings have a difference around one for both Digg and Reddit, which result agrees well with the relationship of $N(t) \sim t^{-c+c_0+1}$ and $\frac{dN(t)}{dt} \sim t^{-c+c_0}$ from our model. One point worth mentioning here is that, in some of existing works about attention dynamics, the above

derived relationship is used as an assumption upon which the model is built [20, 21].

Now, we turn our attention to the distribution of conversation sizes when topics reaches inflection points, i.e. $N(T)$, based on the observed distribution of T. For simplicity, we use $c' = -c + c_0 + 1$, and $\gamma' = \gamma M$ in our derivations, so that $N(T) = \gamma' T^{c'}$. We then have

$$P(N(T) \leq n) = P(\gamma' T^{c'} \leq n) = \mathrm{P}(T \leq (\frac{n}{\gamma'})^{\frac{1}{c'}}). \quad (6)$$

The actual form of the cumulative distribution depends on the distribution of exposure durations, which is website specific as discussed earlier. We now discussed in more detail the impact that different exposure duration distributions have on the empirically measured conversation size distributions. We look at two general cases of exposure duration: (i) exponential distribution and (ii) Pareto distribution.

5.1 Exponential Exposure Duration

For an exponentially distributed T with rate parameter λ as measured in Digg and Reddit, its cumulative distribution has a form of $P(T \leq x) = 1 - e^{-\lambda x}$. Replacing this back to Equation 6, we have

$$\mathrm{P}(N(T) \leq n) = 1 - e^{-\lambda(\frac{n}{\gamma'})^{\frac{1}{c'}}}. \quad (7)$$

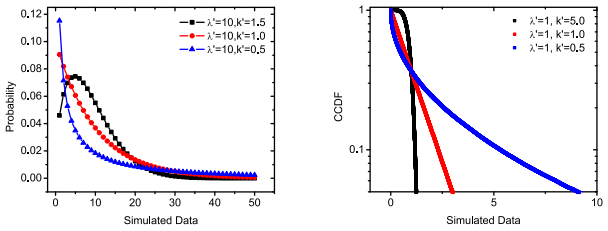
By taking the derivative of this equation, we arrive at the distribution of $N(T)$, which takes the form:

$$\mathrm{P}(N(T) = n) = \frac{\lambda}{c'\gamma'}(\frac{n}{\gamma'})^{\frac{1}{c'}-1} e^{-\lambda(\frac{n}{\gamma'})^{\frac{1}{c'}}}. \quad (8)$$

This is actually a Weibull distribution with its shape parameter k' equals $\frac{1}{c'}$ and scale parameter λ' equals $\frac{\gamma'}{\lambda^{c'}}$. Interestingly, the tail of the distribution scales as $e^{-\lambda(\frac{n}{\gamma'})^{\frac{1}{c'}}}$. So the distribution has following properties:

- **Case 1:** $(k' = \frac{1}{c'} < 1)$ In this case, when the social influence factor has a stronger impact than the decay factor, $c' > 1$, the shape factor is smaller than one. So
$$\lim_{n\to\infty} e^n P(N > n) = \infty, \quad (9)$$
which results in a heavy tailed distribution of conversation size.
- **Case 2:** $(k' = \frac{1}{c'} > 1)$ In this case, the tail decays faster than an exponential distribution. The distribution would appear to be light-tailed.
- **Case 3:** $(k' = \frac{1}{c'} = 1)$ This is the case when the size distribution has an exponential distribution, which is corresponding to the red line in Figure 6 (a) and (b) .

Thus for the case of exponentially distributed exposure duration, both heavy tailed and non-heavy tailed distributions can appear. The actual form of the distributions is determined by the factor c_0. If social propagation dominates, there is a good chance that one would observe extremely large comment threads. Figure 6 (a) demonstrates the simulated density plot under the three cases of heavy-tailed, exponential and light-tailed size distribution. And Figure 6 (b) shows the complementary cumulative distribution function(CCDF) Plot in a semi-log scale of above three cases. If the tail is not exponentially bounded, the CCDF curve will lie above the straight lin,e as seen in the blue one in Figure 6 (b).



(a) Density Plot (b) CCDF Plot

Figure 6: (a) Density plot of simulated size for the situation of exponentially distributed exposure duration, when $\lambda' = 10$ and (i)$k' = 1.5$, (ii)$k' = 1.0$, (iii) $k' = 0.5$. (b) CCDF plot of simulated size, when $\lambda' = 1$ and (i)$k' = 5.0$, (ii)$k' = 1.0$, (iii) $k' = 0.5$.

5.2 Pareto Exposure Duration

Next, we investigate the situation when the exposure duration follows a Pareto distribution, as measured in Epinions. We have

$$P(T < x) = 1 - (\frac{x}{T_{\min}})^{-\alpha}, x > T_{\min}. \quad (10)$$

By taking this back to Equation 6, we have

$$P(N(T) \leq n) = P(\gamma' T^{c'} \leq n) = 1 - T_{\min}^{-\alpha}(\frac{n}{\gamma'})^{-\frac{\alpha}{c'}}. \quad (11)$$

Taking the derivative on n, the size distribution has the form of:

$$P(N(T) = n) \sim \frac{\alpha}{c'}(\frac{n}{\gamma'})^{-\frac{\alpha}{c'}-1}, \quad (12)$$

which is a Pareto distribution. Thus the conversation size of topics with a Pareto exposure duration has a heavy-tail.

From the above analysis of conversation size distribution under different exposure durations, we can see that the discrepancies in the reported size distributions stem from the hidden algorithms that websites employ for deciding which new topics to display on their websites. By separating these artificial factors from user behavior, we show that there are actually underlying mechanisms in common for different social media websites. This explains why different categories of distributions (heavy tailed and non-heavy tailed) are observed in existing studies [1, 7–11]. The model can be adapted to other empirically measured exposure durations. For instance, for a lognormal distribution of exposure duration, a lognormal size distribution is expected from our model. Due to the space limitations, we omit the derivations here. We compare the predictions of this model with empirical measurements in Section 6.

5.3 Structure of Conversation

Another interesting characteristics of on-line conversations is the interactive nature of comments. For example, when a new comment is added to the thread, it is following either the original post or one of the existing comments, so that the comments form a directed graph with each comment as a node. Figure 7 shows one such example based on a sample thread of comments from Digg. The node 0 is the original post of topic, and the others nodes represent the following comments. Node 2 has a in-degree of 2 in this graph. Within

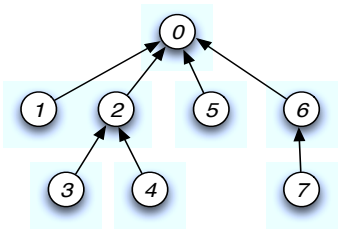


Figure 7: Structure of a sample conversation thread from Digg. Node 0 is the initial post of the topic.

such an information network formed by user comments, one of the most important properties is the in-degree distribution of the associated directed graph. We now show that by adding to our dynamic model for conversations a Yule process [14, 36], one can derive the in-degree distribution of the network. The Yule process assumes that each new comment added into the thread follows one simple rule: it is added to one of the existing comments with a probability proportional to their existing in-degree. Notice that there is also a probability to comment on one with zero in-degree, denoted by δ_0. We use $k_j(t)$ to denote the in-degree at time t for comments that are created at time point j. Based on our model, the number of new comments added into the thread at time t is proportional to $t^{c'-1}$. The cumulative count of comments up to time t is proportional to $t^{c'}$. The sum of in-degree would thus be equal to $(1+\delta_0)\gamma' t^{c'}$. Based on our assumptions on Yule process, the probability of a new comment attaching to an existing comment with k_i in-degree would be $\frac{(k_i(t)+\delta_0)}{(1+\delta_0)\gamma' t^{c'}}$. Since there are $c'\gamma' t^{c'-1}$ new comments added at t, the growth dynamics of $k_j(t)$ can be written as

$$\frac{\partial k_i(t)}{\partial t} = \frac{(k_i(t)+\delta_0)c'\gamma' t^{c'-1}}{(1+\delta_0)\gamma' t^{c'}}. \quad (13)$$

After integrating this equation and taking into account the initial condition $k_i(i) = 0$, we have

$$\ln(\frac{k_i(t)+\delta_0}{\delta_0}) = \frac{c'}{1+\delta_0}\ln(\frac{t}{t_i}). \quad (14)$$

The in-degree of node i at the inflection point equals to

$$k_i(T) = \delta_0[(\frac{T}{t_i})^{\frac{c'}{1+\delta_0}} - 1]. \quad (15)$$

So now we have

$$P(k_i(T) < l) = P(\delta_0[(\frac{T}{t_i})^{\frac{c'}{1+\delta_0}} - 1] < l) = P(t_i > T[1 + \frac{l}{\delta_0}]^{-\frac{1+\delta_0}{c'}}) \quad (16)$$

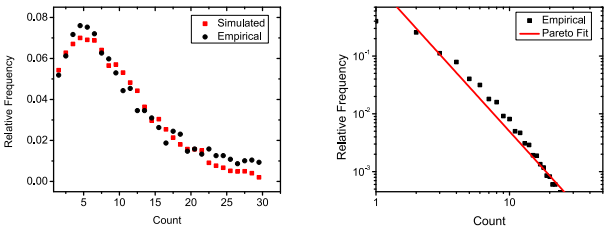
For a randomly chosen node, $P(t_i > t) = 1 - \frac{t^{c'}}{T^{c'}}$, so Equation 16 becomes

$$P(k_i(T) < l) = 1 - [\frac{l}{\delta_0} + 1]^{-1-\delta_0}. \quad (17)$$

By taking the derivative of this equation we see that the distribution of in-degrees scales with $[\frac{l}{\delta_0} + 1]^{-2-\delta_0}$, which amounts to a Pareto scaling. An interesting consequence from the above derivation is that the in-degree distribution within a conversation network is independent of the distribution of exposure durations. That is to say, despite of the different hidden algorithms used by websites, the structure within a conversation thread is universal. This result explains why existing studies observe the same Pareto scaling of in-degree distribution within conversation threads in different social media sites [1, 7–11].

6. EMPIRICAL OBSERVATIONS

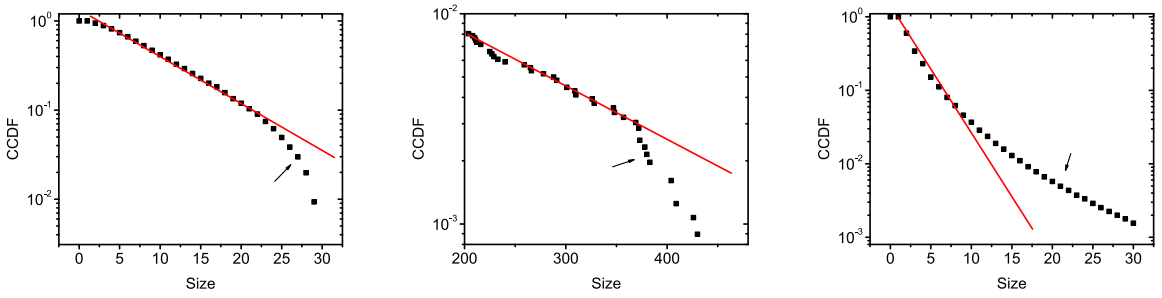
In the last section we modeled the process of conversation growth and predicted that the distribution of conversation sizes is determined by several factors including the exposure duration, the users' commenting behavior, the social propagation and resonating factors. We also demonstrated that a universal Pareto in-degree distribution is expected for each comment. In this section, we compare these results with empirical observations.



(a) Digg (b) Epinions

Figure 9: The density plot of conversation size on (a) Digg and (b) Epinions.

First, we compare the size distribution of Digg conversation size with our model. Since the exposure duration within Digg is observed to follow an exponential duration (Figure 2(a)), and c' is less than one (Figure 5), the size distribution of conversation is expected to be described by the second case in Section 5.1, which is a light tailed Weibull distribution. To exclude the impact of stories promoted to front page in the popular column, we filtered out topics with extreme large comment size. We fit the empirically measured conversation size with Weibull distribution using MLE and estimate that the scale factor equals 10.834 and the shape factor equals 1.439. We use the estimated parameters to simulate the conversation size distribution. The density plot of empirical observation and simulated data is as shown in Figure 9 (a). In Figure 10 (a), we plot the empirically measured size in a Weibull Plot. And in Figure 10 (b), we plot the simulated and empirically measured data in a QQ Plot. The straight lines in both plots demonstrate that the empirically observed conversation size fits well with the predicted size distribution. From the CCDF Plot in semi-log scale as in Figure 8 (a), we can see that the distribution has a light-tail. Similarly on Reddit, the size distribution is expected to follow a lighted-tailed Weibull distribution, as shown in Figure 8 (b). We also measure the size distribution using dataset from Epinions. Based on Figure 2 (b), the size distribution of Epinions is expected to follow a Pareto distribution from results in Section 5.2. The density plot of size in log-log scale is as shown in Figure 9 (b). The Pareto scaling agrees with our model. The distribution is not exponentially bounded as shown in Figure 8 (c). A summary of the tail properties in different social media is shown in Figure 8 (d). The size distributions from Digg and Reddit have a light-tail and that from Epinions a heavy-tail. The



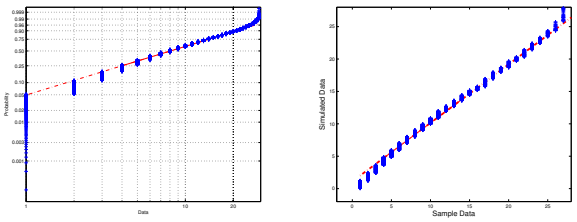
(a) Digg (b) Reddit (c) Epinions

	Digg	Reddit	Epinions
Topic Exposure Duration Distribution	Exponential	Exponential	Pareto
Predicted Size Distribution	Weibull (Light Tailed)	Weibull (Light Tailed)	Pareto
Observed Size Distribution	Weibull (Light Tailed)	Light Tailed	Pareto

(d) Summarization of the size distribution in different websites.

Figure 8: (a) Digg, (b) Reddit and (c) Epinions, the CCDF Plot in semi-log scale for three social media websites. If the density tail (black squares in the plot) is above the red exponential line, then it is a heavy-tailed distribution, otherwise not. (d) Summarization of predicted and measured tail properties in different websites.

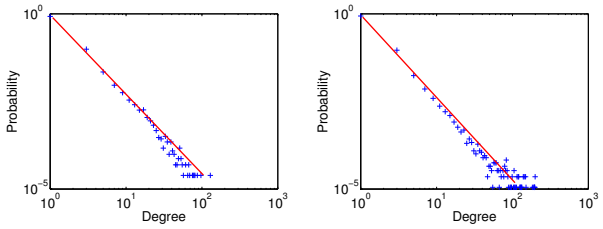
derived size distribution from model agrees well with empirical measurements.



(a) Weibull Plot (b) QQ-Plot

Figure 10: (a) Weibull Plot of empirically observed conversation size. (b) QQ-Plot of empirically measured and simulated conversation size.

In addition, we compare the derived Pareto scaling of in-degree size distribution for different datasets. Figure 11 shows the density plot of in-degree in a log-log scale of Digg and Reddit. We observed the same scaling in Epinions dataset. The straight line in the figure confirms that the in-degree size follows a Pareto distribution. From the above comparisons, our model quantitatively explains the observed discrepancies of size distribution in different social media, as well as the in-degree size distribution in the information network formed by user comments.



(a) Digg (b) Reddit

Figure 11: In-degree size distribution of comments in log-log scale (a) Digg and (b) Reddit. The red straight line in the plots suggests a Pareto distribution.

7. CONCLUSION AND FUTURE WORK

In this paper, we investigated properties of user conversations in on-line social media. We started from the commenting behavior of individual users and the distribution of exposure duration during which the new topics are displayed to users. Based on these observations, we proposed a general dynamic model for conversation growth. The model successfully explains the reported difference in existing studies from different social media websites. We further extended our model with a Yule process to derive the structure of conversations. The results of our model were compared with various empirical measurements, such as the scaling relationship between time and size, the tail properties of size distribution and also the in-degree distribution from different social media sites. Our model provides a powerful framework that can be easily modified and applied to various specific scenarios for studying on-line conversations. We also noticed that for the model to be suitable for the most general cases, we make some simplified assumptions such as the similarities of users and the introduction of inflection point. Possible refinements of the model may take into considerations of the differences between users, the interestingness of topics, and also the impacts of other featuring mechanisms used by the website. In closing, we note that although the focus in this paper has been on user comments and on-line conversations, the framework of our growth model may be suitable to a wide

category of attention dynamics related studies. The wide applicability and the relatively simple assumptions make our model an extremely general one and therefore should provide ample opportunities for future work.

8. ACKNOWLEDGMENTS

C. W. would like to thank HP Labs for financial support.

9. REFERENCES

[1] R. Kumar, M. Mahdian, and M. McGlohon. Dynamics of conversations. In *KDD*, 2010.

[2] Guimer Roger, Uzz Brian, Spiro Jarrett, and Amaral Lus A. Nunes. Team assembly mechanisms determine collaboration network structure and team performance. *Science*, 308(5722), 2005.

[3] Wuchty Stefan, Jones Benjamin, and Uzzi Brian. The increasing dominance of teams in production of knowledge. *Science*, 316(5827), 2007.

[4] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdhury. Twitter power: Tweets as electronic word of mouth. *J. Am. Soc. Inf. Sci.*, 60(11), 2009.

[5] D. Gruhl, R.Guha, R. Kumar, J. Novak, and A. Tomkins. The predictive power of online chatter. In *KDD*, 2005.

[6] M. Kearns, S. Suri, and N. Monfort. An experimental study of the coloring problem on human subject networks. *Science*, 313, 2006.

[7] G. Mishne and N. Glance. Leave a reply: An analysis of weblog comments. In *Third Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics.*, 2006.

[8] V. Gomez, H. Kappen, and A. Kaltenbrunner. Modeling the structure and evolution of discussion cascades. In *HT*, 2011.

[9] V. Gomez, A. Kaltenbrunner, and V. Lopez. Statistical analysis of the social network and discussion threads in slashdot. In *WWW*, 2008.

[10] P. Ogilvie. Modeling blog post comment counts. In *http://livewebir.com/blog/page/2/*, 2008.

[11] M. Tsagkias, W. Weerkamp, and M. de Rijke. Predicting the volume of comments on online news stories. In *CIKM*, 2009.

[12] David Liben-Nowell and Jon Kleinberg. Tracing information flow on a global scale using internet chain-letter data. *Proc. Natl. Acad. Sci.*, 105(12), 2008.

[13] Benjamin Goluba and Matthew O. Jackson. Using selection bias to explain the observed structure of internet diffusions. *Proc. Natl. Acad. Sci.*, 107(24), 2010.

[14] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *KDD*, 2009.

[15] F. Duarte, B. Mattos, A. Bestravras, V. Almedia, and J. Almedia. Traffic characteristics and communication patterns in blogosphere. In *ICWSM*, 2007.

[16] M. De Choudhury, H. Sundaram, A. John, and D. Duncan Seligmann. What makes conversations interesting? themes, participants and consequences of conversations in online social media. In *WWW*, 2009.

[17] B. A. Huberman. The laws of the web: Patterns in the ecology of information. *The MIT Press*, 2001.

[18] Josef Falkinger. Limited attention as a scarce resource in information-rich economies. *Economic Journal*, 118(532), 2008.

[19] Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. Finding high-quality content in social media. In *WSDM*, 2008.

[20] Fang Wu and B. A. Huberman. Novelty and collective attention. *Proc. Natl. Acad. Sci.*, 105(17599), 2007.

[21] Kristina Lerman and Tad Hogg. Using a model of social dynamics to predict popularity of news. In *WWW*, 2010.

[22] R. Crane and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proc. Natl. Acad. Sci.*, 105(15649), 2008.

[23] Jacob Ratkiewicz, Santo Fortunato, Alessandro Flammini, Filippo Menczer, and Alessandro Vespignani. Characterizing and modeling the dynamics of online popularity. *Phys. Rev. Lett.*, 105(158701), 2010.

[24] Sitaram Asur, B. A. Huberman, Gabor Szabo, and Chunyan Wang. Predicting the volume of comments on online news stories. In *ICWSM*, 2011.

[25] A.L. Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435, 2005.

[26] J.P. Eckmann, E. Moses, and D. Sergi. Entropy of dialogues creates coherent structures in e-mail traffic. *Proc. Natl. Acad. Sci.*, 101(14333), 2004.

[27] D. Rybski, S. Buldyrev, S. Havlin, F. Liljeros, and H. Makse. Scaling laws of human interaction activity. *Proc. Natl. Acad. Sci.*, 106(12640), 2009.

[28] B. A. Huberman, P. L. T. Pirolli, J. E. Pitkow, and R. M. Lukose. Social attention in the age of the web. *Science*, 280(95), 1998.

[29] Eytan Adar, Jaime Teevan, and Susan Dumais. Resonance on the web: Web dynamics and revisitation patterns. In *CHI*, 2009.

[30] Anna Chmiel, Kamila Kowalska, and Janusz A. Holyst. Scaling of human behavior during portal browsing. *Phys. Rev. E*, 80(066122), 2010.

[31] S. L. Scott and P. Smyth. The markov modulated poisson process and markov poisson cascade with applications to web traffic data. *Bayesian Statistics*, 7, 2003.

[32] Juliette Stehle, Alain Barrat, and Ginestra Bianconi. Dynamical and bursty interactions in social networks. *Phys. Rev. E*, 81(035101), 2010.

[33] D. Helbing and B. Tilch. A power law for the duration of high-flow states and its interpretation from a heterogeneous traffic flow perspective. *The European Physical Journal B*, 68(4), 2009.

[34] A. Clauset, C.R. Shalizi, and M.E.J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4), 2009.

[35] Erhan Ćinlar. *Introduction to stochastic processes.* Prentice-Hall, Englewood Cliffs, NJ, 1975.

[36] G. Yule. A mathematical theory of evolution based on the conclusions of dr. j.c. willis f.r.s. *Philosophical Transactions of the Royal Society London*, 213(21-87), 1924.