Dynamics of Conversations

Ravi Kumar Mohammad Mahdian Mary McGlohon*
Yahoo! Research
701 First Avenue
Sunnyvale, CA 94089
{ravikumar,mahdian}@yahoo-inc.com

Machine Learning Dept.
Carnegie Mellon University
Pittsburgh, PA 15213
mmcgloho@cs.cmu.edu

ABSTRACT

How do online conversations build? Is there a common model that human communication follows? In this work we explore these questions in detail. We analyze the structure of conversations in three different social datasets, namely, Usenet groups, Yahoo! Groups, and Twitter. We propose a simple mathematical model for the generation of basic conversation structures and then refine this model to take into account the identities of each member of the conversation.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications—*Data mining*

General Terms: Algorithms, Experimentation, Measurement, Theory

Keywords: Conversations, Threads, Graph models, Human response, Usenet, Groups, Twitter

1. INTRODUCTION

In today's world, information networks such blogs, online forums, and other online content-generating communities are among the most important sources of knowledge. In such networks, information is provided and disseminated through social interaction among members of the community. Understanding the dynamics of such interactions is therefore essential in making sense of how this information is generated, how reliable is each piece of information, and how the content generation process can be influenced to achieve better results.

There has been significant research on the dynamics of networks of linked information such as the web, where content providers (webpage authors) form a graph by linking to each other. We know various properties of such graphs, theoretical generative models that provide simple explanations for underlying processes that gives rise to these properties, and methods for using the graph structure in order to extract information about the reliability and importance of various nodes.

*Funded in part by Yahoo! Research

KDD'10, July 25–28, 2010, Washington, DC, USA.

Another important class of information networks that have not received enough attention from a theoretical point of view are those where units of information have relatively short life-spans (shorter than that of webpages), and therefore time is an essential part of the dynamics of network creation. This class includes Twitter, online discussion forums, and news websites. With the modern-day shortened news cycle, such networks are becoming increasingly important.

In this paper, we seek to study a class of such networks, namely, the network formed by conversation threads in online communities. We will examine three different data samples: Yahoo! Groups, Usenet, and Twitter. In addition to the important role time plays in the growth of such networks, a factor that makes them a particularly desirable object of study is that the links in these networks have a more or less uniform meaning: a link from a node u to another v means that the message corresponding to u is in reply to that of v. Each node is in response to only one other node (if any), and furthermore, each node is identified with a distinct author. In other words, nodes in these networks are more atomic units of information than, say, webpages.

Our goal is to shed light on how conversations form in different online groups by studying several questions: What are the common properties of conversation threads? What similarities and differences can be observed between different groups? Can we build models to capture these properties? Can we analyze them? How can we characterize group conversations (i.e., conversations that engage a group of users) as opposed to those that are primarily pairwise exchanges?

The rest of this paper is organized as follows: We start by reviewing the related work in Section 2. In Section 3, we describe the three main datasets that we use in this study: Yahoo! Groups, Usenet, and Twitter. We will first examine the properties of the threads and the induced social networks in Section 4. In addition to the usual observations of the power-law/heavy-tailed distributions, we make two key observations by analyzing the data:

(1) the depth of such threads grow sub-linearly but superlogarithmically in the size of the thread, and

(2) a law similar to Heaps' law [15] holds for the number of distinct authors in a thread.

Section 5 and Section 6 contain a description of generative models for conversation threads. For each model we prove a number of theoretical results, show simulation results, and give results of learning algorithms on our datasets on the model (to learn the parameters of the model). We start with the branching process model, which is the clas-

sical model in probability theory for generating trees (similar to what Erdös–Rényi model is for generating graphs); this serves as a baseline model. Next, in Section 6 we give a preferential-attachment-type model [1] that combines the rich-gets-richer principle with the element of time. We also give a model for the distribution of the authors of the messages in a thread that is based on a variant of the copying process [18] in Section 6.3. Finally, we give a mixture model for forums such as Twitter where the types of threads we observe are not homogeneous, and show how expectation-maximization algorithms can be used to partition the conversations into different classes. Section 8 contains some anecdotal examples from our dataset, based on our study. Section 9 contains concluding remarks.

2. RELATED WORK

The related work falls into the following main categories: work on conversations and human activity in general, cross-community group dynamics, and graph models.

Conversations and human activity. The work closest to ours is that of Liben-Nowell and Kleinberg, who studied the structure of chain letter propagation [25], showed that the structure was characterized by a deep tree-like pattern, and proposed a probabilistic model to generate such trees. Golub and Jackson [12] built on this to show that a basic branching process model combined with the selection bias of observing only large diffusion can explain the results in [25]. These lines of work concern a mechanism for the spread of information in a social network. We, on the other hand, are interested in studying the patterns of interactions and repeated interactions (i.e., conversations) in closed groups.

Another work related to ours is that of Leskovec, Backstrom, and Kleinberg [20], who considered the propagation of "memes" across the Web in the context of news cycle. In course of studying this problem, they consider a model where they combine recency and the preferential attachment process. However, their focus is not on a graph generation model and, as they indicate, the combining form they propose does not seem to be analysis-friendly.

There has also been some exploration into the dynamic processes of conversation and information propagation. Barabasi [3] postulated that the bursty nature of human behavior is a consequence of a decision-based queuing process and used it to explain the heavy-tailed activity patterns in e-mail communications; Vazquez et al. [8, 31] further explored this model. This was reproduced in [23], where response times to blog posts were shown to have a similar heavy-tailed distribution.

Conversations can also be characterized as *information cascades*, phenomena in which an action or idea becomes adopted due to the influence of others, typically, neighbors in some network [4, 11, 13]. Cascades on random graphs using a threshold model have been theoretically analyzed [33]. There has been empirical analysis of the topological patterns of cascades in various contexts, such as recommendation networks [24, 19] and blog posts [23]. In the latter, authors extracted "cascades", or conversation threads, from a large set of blog posts, and studied patterns with respect to the sizes and shapes of these cascades, as well as topological aspects of the network at large. They continued this to show that different genres of blogs have different patterns of cascade shapes [27].

Cross-community studies. There have been several previous studies across social networks data. Backstrom et al. studied Yahoo! Groups data, defining "thriving" groups and tracking engagement of core users in groups [2]; see also [7]. Kumar, Novak, and Tomkins studied the topological structure and component size distribution of Flickr and Yahoo! 360 networks, identifying "star" structures and showing how they persisted and eventually joined the giant component [17]. Leskovec et al. studied the edge arrivals of different online networks, proposing a generative model [21].

There has been a significant body of work on forum data. Microsoft's Netscan Project has conducted a very thorough study of Usenet discussion patterns, depicting hierarchy of newsgroups and their changes between 2000 and 2004 [30], studying the *social roles* of Usenet authors [10], and creating a visualization tool for different author roles identified [32]. Other authors explored the network structure of different groups and studied cross-posting behavior [26].

Graph models. There has been a lot of work on developing tractable mathematical models for real-world graphs and social networks, starting with the legendary Erdös–Rényi G_{np} model. For a detailed survey of these models, the readers are referred to [6, 9, 16]. There have been a few developments on graph models since these surveys, e.g., [22, 21]; these are beyond the scope of our work. To the best of our knowledge, group conversations have not been explicitly addressed in any of the previous works.

For a detailed background on branching processes, the readers are referred to the classic book by Harris [14].

3. PRELIMINARIES

3.1 Data description

We first describe the three sources of data that will be used in our study, namely, messages from a set of Usenet groups, messages from a set of public Yahoo! Groups, and Twitter tweets over a month. The first and the last datasets are publicly available and hence our experiments and observations are repeatable.

Each dataset consists of records, where each record has the ID of message, the ID of its parent message (if applicable), the author of the message, and a timestamp. Notice that all the three datasets enable conversations among users, i.e., messages can be posted in response to earlier messages. We will use these sources to show some commonalities in thread structures.

Usenet groups. Usenet is a decentralized set of forums across different subjects and languages. We sampled Usenet based on groups posted to in early January 2010, according to `http://newsadmin.com/top100tmsgs.asp`, using the server Giganews. For a complete list of the groups crawled, refer to `http://www.cs.cmu.edu/~mmcgloho/pubs/groupthreads-list.txt`. This gave us a broad sample of newsgroups, including some on political discussion (`alt.politics`, `it.politica`), recreational activities and hobbies (`rec.outdoors.rv-travel`, `rec.music.beatles`), and general news or ads (`news.lists.filters`, `alt.marketplace.online.ebay`). This crawl produced around 10 million posts in total. Most groups had between 1,000 and 5,000 users, with some as few as 20. We also had a deeper crawl that focused only on political

Dataset	Messages ($\times 10^6$)	Threads ($\times 10^6$)	Users ($\times 10^6$)
USENET	22.61	3.896	1.659
Y!GROUPS	5.869	1.558	0.690
TWITTER	69.94	36.24	5.023

Table 1: Synopsis of the datasets.

groups, see [26]. This consisted of around 200 groups with posts from 2004–2008. This included several general politics groups (`alt.politics`, `talk.politics.misc`), some national politics groups (`it.politica`, `uk.politics`), state or regional groups (`pa.politics`, `bc.politics`), and topical groups (`uk.politics.guns`, `talk.politics.drugs`). This produced 37 million posts.

While Usenet is declining in popularity, it has the feature that it is public, easy to crawl, and has an obvious thread structure (with reply-to as a line in the header). Furthermore, certain Usenet groups are still very active, and have not declined in usage. This is the rationale for having Usenet as part of our data sample.

Yahoo! Groups. Yahoo! Groups is a popular online groups application. We chose public groups from Yahoo! that were moderated (unmoderated groups were mostly spam); we restricted our attention to groups that were still active, i.e., they were not deleted or suspended. We also restricted the sample to groups that had at least ten messages and had at least ten distinct users. This resulted in 13,102 groups in the dataset with over 14.9 million posts. The groups in our data included ones such as `WrestlingGear`, `cookbook-reviews`, `IndianaSPCA`, `welcometomorocco`, `neurosurgeonsclub`, etc. These groups covered a broad set of topics and interests. Most groups contain 500 to 5,000 users, with some as few as ten (our minimum threshold for including in the dataset). The data was collected in January 2010.

Twitter. Twitter is an extremely popular social application where users send short messages (called *tweets*), sometimes in response to other messages. We examined a large subset of tweets for the month of September 2009. Since the tweets are small (at most 140 characters), in addition to the message meta-data (which includes reply-to information), we have the entire message itself! This allows us to use the message content for our study, if needed.

For each of these datasets, we first ran an algorithm to find the threads, which in this case are the connected components. This partitioned the data into threads, forming the basis of our study. Table 1 gives a high-level view of the datasets.

3.2 Notation

We use the following conventions in our paper. We denote messages by letters $u, v, w, \ldots$. Messages are assumed to have a thread structure, i.e., each message v is either a new message or is a message in response to an earlier message u. In the latter case, we call u to be the *parent* of v (denoted parent(v)) and v to be a *child* of u. A message with no children is a *leaf* message and a message with no parent is the *root* message. Thus, the root message, along with its descendants form a connected component (in particular, a rooted tree), which we call a *thread*. All the messages in a group can be decomposed into disjoint threads. For a given thread and a message u, let path(u) denote the set of messages from u to the root of the thread.

Each message u has a *timestamp* $t(u)$ associated with it. The messages in a thread are created chronologically and hence if u is a parent of v, then $t(u) \leq t(v)$. The *author* $a(u)$ of a message u is the person who wrote it. A single person can author multiple messages in a thread. Let A be the set of all authors; $a \in_U A$ denotes that a is chosen uniformly at random from A.

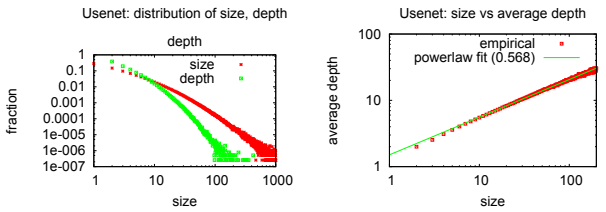
4. PROPERTIES OF CONVERSATIONS

In this section we state the main observations about the threads from our three datasets. The observations we make here are the basis behind the development of our generative models.

Most of the observations are illustrated for USENET; the other two datasets follow mostly similar qualitative patterns, although the actual parameters vary, and are omitted for space.

4.1 Size and depth

We study the distribution of thread sizes and depth (which is the length of the maximum path to a leaf from the root in a thread). Figure 1(a) shows the size and the depth distribution in USENET. As we note, not surprisingly, these are both heavy-tailed.



(a) (b)

Figure 1: (a) Size and depth distributions and (b) size vs depth in USENET.

Next, we consider the relationship between size and depth: what is the average depth of a thread of a given size? Figure 3(b) plots this data. It somewhat surprising that there is a power law relationship between size and depth — the size is roughly quadratic in depth. This observations hints that traditional models such as preferential attachment are provably insufficient to model conversation threads, since such models generate graphs with logarithmic diameter.

4.2 Degree

We next study the degree distribution p of the threads. The degree distribution for USENET is shown in Figure 2. From Figure 2, it is arguable that the degree distribution is close to a power law, i.e., $p(k) \propto k^{-\alpha}$ for some $\alpha > 2$.

Let $\mu = \mathrm{E}[p]$, the mean of the distribution p. Values of μ and α for the three datasets are shown in Table 2.

Next we ask the question: is the degree distribution independent of the level of a thread? Figure 3 shows the degree distribution at each level of the thread (the root is assumed to be at level 1). It is easy to see that the distribution be-

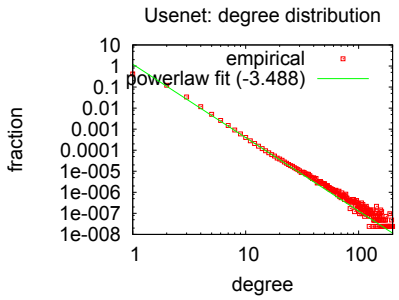


Figure 2: Degree distribution of threads in USENET.

Dataset	μ	α
USENET	0.906	3.488
Y!GROUPS	0.762	2.302
TWITTER	0.657	2.260

Table 2: Values of $\mu = \mathrm{E}[p]$ and α.

comes "steeper" with the level since having more children becomes less likely at higher levels.

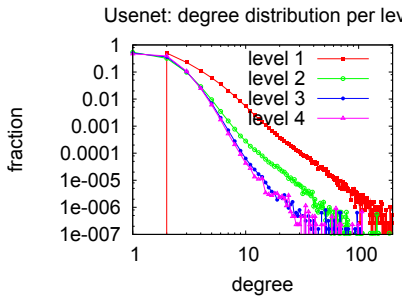


Figure 3: Per-level degree distribution in USENET.

4.3 Authorships

We study the properties of authors of messages in a thread. We first consider the size of a thread and the average number of distinct authors in the thread. We also consider the average of the most number of times an author occurs in a thread. Figure 4 shows these plots. We find that there is a polynomial relationship between the size of a thread and the number of authors participating in the thread. In fact, this relationship is very reminiscent of the *Heap's Law* in information retrieval [15], which relates the vocabulary size to the document collection size.

5. BRANCHING PROCESSES

The Galton–Watson branching process is a classical model in probability theory for generating a random tree. This models many phenomena like the growth of a population (birth processes), and are important objects in random graph theory [5]. In this section we study this model as a generative model for threads, and discuss properties of the real conversations that they do or do not satisfy. This is perhaps the most basic tree generation model, and serves

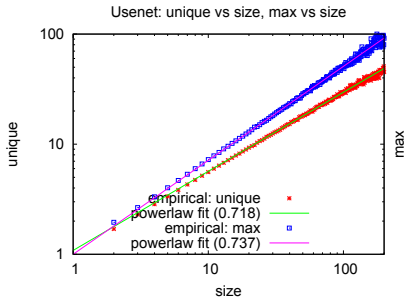


Figure 4: Average number of unique authors and maximum author activity vs thread size in USENET.

as a benchmark for us, similar to the role the Erdös–Rényi $G_{n,p}$ model plays in graph generation.

Recall that in branching processes, each individual in generation i produces a random number of individuals in generation $i+1$ according to a probability distribution. These random numbers are drawn independently for different nodes.

5.1 The branching process model: BP-MODEL

Let p be a fixed probability distribution on non-negative integers. The messages in a thread are generated by the following process. Each thread starts with a root node and proceeds in discrete steps. At the ith step of the process, each leaf at the ith level of the thread constructed so far independently generates a certain number of children according to the distribution p, i.e., a leaf u has k children with probability $p(k)$. If $k = 0$, then u is a leaf. If $k > 0$, then the children of u participate in the $(i+1)$st step. The process terminates when there are no more new children.

Notice that the only parameter of the model is the distribution p. We can fit the real dataset to BP-MODEL and compute the maximum likelihood estimate for this parameter: $p(k)$ is estimated to be the fraction of nodes with k children in the data; it can be easily shown that this is indeed the maximum-likelihood estimator. BP-MODEL can simulate the inferred distribution in order to generate the threads.

5.2 Properties of BP-MODEL

Let Z_i be the random variable denoting the number of children at the ith level of the threads. Let $Z = \sum_i Z_i$ be the random variable denoting the size of the thread. From the definition of a branching process, the mean size of a thread is given by the recurrence

$$E[Z] = 1 + \sum_{j=1}^{\infty} ip(i)E[Z] \Longrightarrow E[Z] = (1-\mu)^{-1}.$$

In our case, from Table 2, since $\mu < 1$ for all three datasets, the branching process dies out almost surely.

We now analyze the tails of two properties of the threads generated by the model, namely, their size and their depth. We first show that the tail of the size distribution is quantitatively similar to that of the degree distribution. Let $X \sim p$ be a random variable distributed according to p.

LEMMA 1. *For any $i > 0$ and $k > 0$, $E[X^k] < \infty$ if and only if $E[Z_i^k] < \infty$.*

PROOF SKETCH. It is easy to see that the size distribution stochastically dominates the degree distribution. Therefore, if the degree distribution does not have a finite kth moment, then the size distribution also does not have a finite kth moment.

Conversely, we show that if the degree distribution has a finite kth moment, then the kth moment of the size distribution is also finite. For simplicity, we illustrate this for $k = 2$. From the basic theory of branching processes [14], the generating function for Z_i is given by the ith iterate f_i of the generating function f of p. The second moment of Z_i is given by $f_i''(1)$. We know that $f_1'(1) = f'(1) = \mu$ and let $f_1''(1) = f''(1) = \nu < \infty$ by assumption. It is also easy to see that $f_i''(1) = \mu^i$. By simple calculations, one can obtain the recurrence

$$f_i''(1) = f'(1)f_{i-1}''(1)+f''(1)(f_{i-1}'(1))^2 = \mu f_{i-1}''(1)+\nu\mu^{2(i-1)},$$

from which

$$f_i''(1) = i\nu\mu^{i+1}\frac{\mu^i - 1}{\mu - 1} < \infty.$$

□

An important corollary of the above lemma is that the distribution of the size of a the tree generated using a branching process follows a heavy-tail distribution[1] if and only if the distribution of the number of children is heavy tailed.

Next, we analyze the depth of threads generated by the model. We show that the depth has an exponential vanishing tail.

LEMMA 2. *If $\mu < 0$, the probability that the tree generated by the branching process has depth at least i is exponentially small in i.*

PROOF. The expected number of children in the ith generation is given by $\mathrm{E}[Z_i] = \mu^i$. For a tree to have depth at least i, this number must be at least 1. By the Markov inequality, the probability of this event is at most $\Pr[Z_i \geq 1] \leq \mathrm{E}[Z_i] = \mu^i$. □

From this, we see that the distribution of depths of threads generated by BP-MODEL does not in particular have a heavy tail.

5.3 A critique of BP-MODEL

The main advantage of BP-MODEL is its conceptual simplicity. Furthermore, it is also easy to estimate the parameters of the model, and as we observed, the parameter (i.e., the degree distribution) can be succinctly approximated by a power law. By Lemma 1, it also leads to a heavy-tailed size distribution, provided the degree distribution is heavy-tailed (see Figure 1). As we will see in Section 7, BP-MODEL is sufficient to elicit different types of conversations.

The main drawbacks of BP-MODEL are the following.

(1) The model is not generative, i.e., the degree distribution is stipulated and the messages are created according to this distribution. In this sense, this model is similar to the configuration model [28] in random graph theory, where a random graph with a specified degree sequence is generated. The model does not try to abstract the social processes behind the creation of messages and the growth of threads.

[1]By a heavy-tail distribution we mean a distribution that dominates a power law distribution for some exponent.

(2) This model cannot capture the depth distributions of threads that are observed in reality (Figure 1(a)). From Lemma 2, we know that the depth cannot be heavy-tailed; this is seen in the in Figure 5. BP-MODEL also cannot capture the quadratic relationship between size and depth in Figure 1(b).

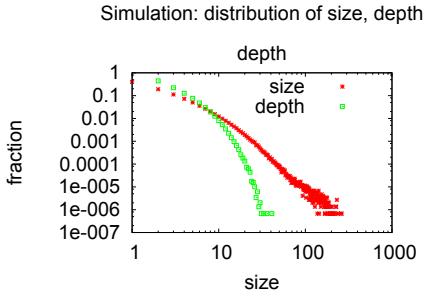


Figure 5: Size and depth distribution of threads using BP-MODEL (with p estimated from USENET).

Moreover, the size distribution generated by the model has a tail that is quantitatively similar to that of the degree distribution. However, in reality, the size distribution has a flatter tail than the degree distribution (for sake of brevity, we do not show these figures).

(3) In the branching process model, the number of children at each node is determined by a single distribution. However, this is not realistic as seen in Figure 3. A vanilla branching process model cannot capture this phenomenon.

(4) The branching process model does not capture the order in which the messages are created, i.e., the timestamps associated with the messages are left out. Furthermore, the model does not capture the author of messages. These are two critical parameters that distill the essence of conversations in social settings.

6. MODELS WITH TIME AND IDENTITY

In this section, we propose new models for the growth of threads of conversation.

First, we consider a model that incorporates recency. The idea behind this model (called T-MODEL) is based on the following observation: as in the preferential attachment model, messages that have already received many replies are more likely to receive a new reply. But in addition to this, new messages receive more attention than the old ones. This effect might not be very pronounced in the growth of networks such as the web where the nodes (webpages) have a relatively long "lifespan". On discussion forums and Twitter, however, messages quickly become outdated, and therefore (as we will demonstrate later in the paper using data and simulation) there is a clearly observable tendency that a new message added to a thread is in response to a relatively recent message. Our model captures this fact by assigning a higher probability of being the next message that receives a reply not only to high-degree messages, but also to the recent messages.

As noted in Section 2, a similar high-level idea was explored in the context of tracking news phrases [20].

6.1 Definition of T-Model

We now give a formal definition of T-Model. We assume the thread grows in discrete time steps. Each time, either a decision is made to stop the thread (i.e., no more message will be added to it), or to add a message in reply to one of the current messages in the thread denoted by v (i.e., the new node will be added as a child of v). The probability of the latter decision depends on two parameters of the node v. One parameter is the current degree of v; we denote this by $\deg_v$. The other parameter, called the *recency* of v and denoted by r_v, is the number of time steps since v was added to the thread.

In general, we take the probability of the decision to add a child to v to be proportional to some function $h(\deg_v, r_v)$ of the degree and recency of v, and the probability of death to be proportional to a constant δ. That is, the probability of adding a child to v is $\frac{h(\deg_v, r_v)}{\sum_u h(\deg_u, r_u)+\delta}$ and the probability of termination is $\frac{\delta}{\sum_u h(\deg_u, r_u)+\delta}$, where the summation in the denominator is over all nodes u currently in the thread.

For the rest of this paper, we focus on a particular form of the function h: when h is a linear combination of $\deg_v$ and and an exponentially decreasing function in r_v. That is, $h(\deg_v, r_v) = \alpha\deg_v + \tau^{r_v}$ for constants $\alpha \geq 0$ and $\tau \in (0, 1)$. We choose this form of function because of the following:

(1) An exponential "discounting" function like τ^{r_v} is the standard way to model dependence on time.

(2) A linear combination is perhaps the simplest and most natural way to combine the recency and the degree[2].

(3) Considering a linear combination (as opposed to, e.g., the square root of the degree plus the exponential discount) allows us to compute the denominator of the probability expressions independent of the current degrees, and this makes this model particularly amenable to mathematical analysis, as we see in this Section.[3]

Note that both the degree and recency components play a role in generating different types of threads. If the former plays a prominent role, then we get "bushy" threads — where many messages are in response to a single earlier message. If the latter plays a prominent role, then we get "skinny" threads — where the thread is essentially a path and messages appear in succession as a cascade of responses.

6.2 Properties of T-Model

In this section we show that the degree distribution of graphs generated from T-Model has a heavy tail.

Theorem 3. *Let G be a thread with n nodes generated from the model in the above section with $h(\deg_v, r_v) = \alpha\deg_v + \tau^{r_v}$. Then for every d, the fraction of nodes of G that have at least d children is at least $\Omega(d^{-1})$.*

Proof Sketch. With $h(\deg_v, r_v) = \alpha\deg_v + \tau^{r_v}$, at the time that the thread has k nodes, we have

$$\sum_u h(\deg_u, r_u) = \alpha(k-1) + \sum_{j=1}^{k} \tau^j < \alpha(2k-2) + \frac{\tau}{(1-\tau}.$$

Now, we consider the ith node added to the thread, and study the growth of the degree of this node at time t, as t grows. We denote the degree of this node at time t by $d_i(t)$. Note that $d_i(t)$ is a random variable, and $d_i(t+1) - d_i(t)$ is either one (if the $(t+1)$'st node connects to i) or zero (if it doesn't). The probability that $d_i(t+1) - d_i(t) = 1$ is

$$\frac{h(\deg_v, r_v)}{\sum_u h(\deg_u, r_u) + \delta} = \frac{\alpha d_i(t) + \tau^{t+1-i}}{\sum_u h(\deg_u, r_u) + \delta} > \frac{\alpha d_i(t)}{\alpha t + \tau/(1-\tau)}.$$

Therefore, we have

$$\mathrm{E}[d_i(i+1)] \geq 1 \tag{1}$$

and

$$\mathrm{E}[d_i(t+1)] - \mathrm{E}[d_i(t)] > \frac{\alpha \mathrm{E}[d_i(t)]}{\alpha t + \tau/(1-\tau)}. \tag{2}$$

We couple the sequence of random variables $d_i(i+1), d_i(i+2), \ldots$ with another sequence which instead of the inequalities (1) and (2), satisfies the corresponding equalities. We call these random variables $d'_i(t)$. By coupling, $d_i(t)$ stochastically dominates $d'_i(t)$. Therefore, it is enough to prove the desired lower bounds on $d'_i(t)$ instead of $d_i(t)$. To do this, we first calculate the expected value of $d'_i(t)$, which we denote by $ED_i(t)$. This can be calculated from the recurrence relations given by (1) and (2). The solution of these recurrences is

$$ED_i(t) = \frac{\alpha t + \tau/(1-\tau)}{\alpha(i+1) + \tau/(1-\tau)}.$$

The above equation can be proved easily by induction on t using recurrences given by (1) and (2). This means that for every i, the expected degree of the ith node of the thread grows at least linearly with time. Furthermore, the sequence of random variables $d'_i(t)$ defines a martingale, and therefore by standard martingale concentration inequalities [29], if $t-i$ is large enough, the value of $d'_i(t)$ is concentrated around its expectation. Putting these together, we obtain that for $t = n$ large enough and $i < n - O(1)$, with a large probability, we have

$$d_i(n) > \frac{\alpha t}{2(\alpha i + \tau/(1-\tau))}.$$

This means that the number of nodes that have degree at least d is bounded from below by the number of i's satisfying $\alpha i + \tau/(1-\tau) < 0.5\alpha n/d$, which is $\Theta(n/d)$. Thus, the fraction of nodes having degree at least d is at least $\Theta(d^{-1})$ □

6.3 Modeling author identity: TI-Model

Upon understanding the process from which the thread structures are generated, we may also want to understand *who* is responsible for generating the reply message.

In this section we propose a model (called TI-Model) for author identity. The motivation for TI-Model comes from the observation that authors tend to respond to responses to their own earlier messages. Thus, when a new message v arrives as a child of message u in a thread t, the author $a(v)$

[2]Another natural alternative that we considered is the product of the degree with the exponential discounting term, i.e., $h(\deg_v, r_v) = \tau^{r_v}\deg_v$. While this formulation might makes sense intuitively, it does not generate graphs similar to what we see in practice. In particular, the exponential discounting factor does not let the degrees of the nodes to grow to a heavy-tailed distribution.

[3]We have also done simulations with a few other reasonable choices of h, and did not observe fundamentally different results.

Dataset	α	τ	δ
USENET	0.1	0.94	0.4
Y!GROUPS	0.7	0.95	0.8
TWITTER	0.1	0.90	0.8

Table 3: Parameters of T-MODEL.

is likely to be chosen from the set $\{a(w)\}$ for some w along the path from u to root(t). (There is a slight caveat that w is unlikely to be u since $a(v)$ is most likely not the same as $a(u)$.)

The above observations, combined with the empirical evidence of Heap's law (Figure 4), suggests a modified Polya urn process in order to reproduce author identity patterns. When a new message v arrives with $u = \text{parent}(v)$, then $a(v)$ is chosen according to the following process. Let $A'(v) = \text{path}(\text{parent}(v))$.

$$a(v) = \begin{cases} a(w), w \in_U A'(v) & \text{wp. } \gamma \\ u & \text{wp. } \epsilon \\ a \in_U A & \text{wp. } 1-\gamma-\epsilon \end{cases}$$

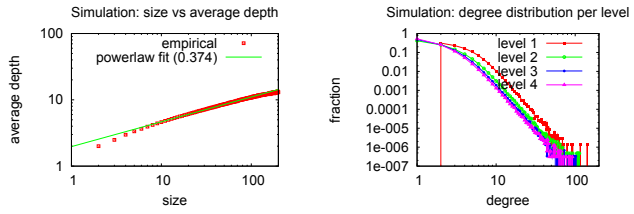
Note that this model can also be viewed as a variant of the *copying* model [18]: with probability $\gamma > 0$, we copy one of the authors from path(parent(u)); with probability $\epsilon \ll \min(\gamma, 1-\gamma)$, we copy u itself; and with the remaining probability, we choose a random author from A. By this process, the probability that an author is chosen is proportional to the number of times he/she already authored a message in the path to the root.

From data, it is easy to statistically learn the parameters γ and ϵ of TI-MODEL. It is possible to show that the above modified Polya urn process generates a heavy tail for the number of occurrences of an author on a path (proof omitted). However, it seems much harder to analyze the number of occurrences in a tree, since different paths share nodes.

6.4 Simulation of the models

In this section we estimate the parameters of TI-MODEL from the data and simulate the model to see if the statistics match the empirical findings. The parameters are estimated through a simple grid search and maximum likelihood computation. Table 3 shows the parameters of T-MODEL estimated from the data.

We consider the size vs depth relationship and the degree distribution conditioned at each level, to see if these resemble the empirical observations. Figure 6 shows these plots for USENET, simulated using the parameters from Table 3. These show that T-MODEL is able to reasonably capture the empirical observations.



(a) (b)

Figure 6: (a) Size vs depth (b) Per-level degree distribution for T-MODEL simulation of USENET.

Finally, we consider the number of unique authors as a function of thread size, by using TI-MODEL. Figure 7 shows the plot. We can see that this is reasonably consistent with the observation we made in Figure 4.

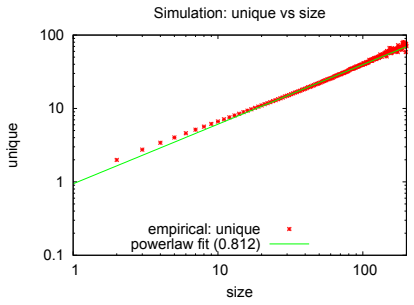


Figure 7: Unique authors vs thread size in TI-MODEL.

7. MIXTURE MODELS AND AN EM ALGORITHM

Communities such as Twitter, Yahoo! Groups, and Usenet are quite diverse, and as a result, sometimes we observe threads with very different characteristics on these communities. In particular, one can observe that on twitter, threads of conversation are primarily of two different types: conversations that are mainly between two individuals, and conversations that are among a group of individuals. For the former type of conversation, the thread is "skinny", growing more or less as a path (sometimes with few additional leaves), whereas for the second type, the thread is often "bushy." To more accurately model the threads in such settings where there is heterogeneity in the types of threads, we consider models that are *mixtures* of the models proposed in previous sections. In particular, a simple model is to consider mixtures of BP-MODEL, with different parameters: each thread is of one of the types $1, \ldots, k$, where the probability of each type is given. Given the type τ, the thread is generated according to a branching process with probability distribution $p^{(\tau)}$ for the number of children of each node.

A useful application of a mixture model is that by fitting the data to such a model (i.e., estimating the maximum likelihood parameters of the model) we obtain a classification of the threads in the dataset. For example, we have applied the method on TWITTER with $k = 2$, and the resulting clustering of the thread matches the intuitive clustering between the long threads of pairwise conversations and the wider group conversation threads. Figure 8 shows the values of the parameters BP-MODEL (i.e., the degree distribution) for TWITTER for $k = 2$. Clearly, $p^{(1)}$ corresponds to the bushy threads and $p^{(2)}$ corresponds to the skinny threads.

To fit the data to a mixture model, we use an adaptation of the well-known *expectation-maximization* (EM) algorithm. The EM algorithm starts with a random partitioning of the threads into k classes. In each iteration, for each class, the algorithm estimates the maximum likelihood set of parameters (in the case of branching processes, this is simply counting the number of nodes with a certain number of children). Then, fixing these sets of parameters, the

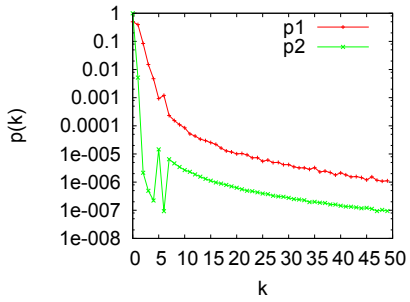


Figure 8: Values of $p^{(1)}, p^{(2)}$ for TWITTER and BP-MODEL.

algorithm reclassifies each thread to the model that is most likely to have generated it. This algorithm continues until it converges, or a certain maximum number of iterations is reached.

We state a few observations about the EM algorithm applied to our datasets (in particular the twitter dataset, which appears to be the most heterogeneous among the three): first, the algorithm converges quite fast. The median number of rounds it takes for the algorithm to converge is 11. This is significant given the size of the dataset.

In fact, it is not hard to prove that the EM algorithm as described above always converges in a finite number of rounds. This is based on the fact that the likelihood of the current calculated solution in this algorithm always increases. However, the convergence is to a local maximum of the likelihood function, and not necessarily a global maximum. In our experiments, the difference between the log-likelihoods of the solution in 10 different runs of the algorithm is less than 0.01%. Furthermore, the parameters calculated for the classification in different runs are almost equal.

8. ANECDOTAL EXAMPLES

We next examine some of the groups in particular for those with the highest values for α (high degree of preferential attachment), τ (high recency effect), and low/high values of γ (high/low copying effect).

8.1 USENET

Preferential behavior. The USENET groups with the highest degree of preferential attachment are shown below. Nearly all of the top ones were politically-related. There were a few additional high-activity non-political groups with somewhat higher values (e.g. `rec.games.pinball` had a value of 0.7), but this is significantly less than those shown.

Group	α
`it.discussioni.leggende.metropolitane`	10
`it.politica.polo`	10
`rec.games.chess.politics`	3
`bln.politik.rassismus`	2
`sk.politics`	1.5

This would lead us to believe that political groups tend to have "bushier" threads, and less "back and forth" paths between a few people.

Recency. On the other hand, there are some groups that had a higher recency effect — some of the lower traffic politics groups (those with fewer users overall) tended to follow this pattern. The top groups are shown below.

Group	τ
`fa.linux.kernel`	0.98
`uk.politics.electoral`	0.98
`rec.arts.drwho`	0.97
`uk.politics.crime`	0.97
`chile.soc.politica`	0.96

Identity "copying". Finally, we examined which groups had the highest and lowest rates of identity copying; that is, which groups showed the highest incidence of choosing authors from upwards in the thread (as opposed to more uniformly distributed). High values of γ indicate a low copying rate — new authors tended to join in often. Low values of γ indicate a low copying rate. Here are some of the higher and lower γ values among USENET— there were 13 total groups with the highest γ; we show a selection.

Group	
`or.politics`	high γ
`alt.fan.cecil-adams`	
`alt.marketplace.online.ebay`	
`pl.misc.kolej`	
`rec.arts.sf.written`	
`linux.debian.bugs.dist`	low γ
`microsoft.public.excel.misc`	
`microsoft.public.excel.programming`	
`nctu.talk`	
`tw.bbs.campus.nctu`	

Interestingly, nearly all of the `rec.music` groups followed a pattern of low copying, with more uniform behavior. The highest copying groups were IT-help related groups, which is not surprising given the back-and-forth question/answer format that such groups foster.

8.2 Y!GROUPS

We repeated the experiments of determining α and τ for the Y!GROUPS data. While the characterization of the groups was less obvious, we show a few groups with unusually high values of each parameter ($\alpha = 10$ and $\tau = 0.99$.)

Group	
`indianmedical`	$\alpha = 10$
`IllinoisSpeakers`	
`DetectiveRichardHead`	
`Bodybuildersvsaverageguys`	
`villageDesign`	
`NorthCarolinaSpeakers`	$\tau = 0.99$
`stbaseliosorthodoxchurch`	
`LostnFoundEvents`	
`PatriceVinci`	
`molecular-biology-notebook`	

8.3 TWITTER

Finally, we repeated the experiments for TWITTER. To find out the topic of a thread, we chose the most popular hashtag (`#tag`) among the messages in a thread and assume it to denote the topic of the thread.

The topics with the highest α and the topics with the highest τ are shown next.

Tag	
#mustsee	$\alpha = 10$
#twitterinreallife	
#readingrainbow	
#whathappenswhen	
#vogueevolution	
#yankees	$\tau = 0.99$
#warriors	
#tiff09	
#mustsee	
#iranelection	
#followfriday	

The first set corresponds to topics with "bushy" threads and the second set corresponds to topics with a stronger sense of time (sports, movies, etc.) and hence the threads tend to be "skinny."

9. CONCLUSIONS AND FUTURE WORK

In this paper we studied the problem of how online conversations build. We proposed simple mathematical models that can capture the patterns in human exchanges. Our models encapsulate both time of the message and identity of the author of the message. Using three different publicly available datasets, we study the structure of conversations and explore the model for these datasets.

There are several potential future applications of this work. We identify two such applications

(1) Our method can be used to identify group conversations in systems like e-mail, and offering tools to facilitate such conversations.

(2) Our method can be used to identifying friendship relationships between users: declaring someone a friend or following someone on social applications like Facebook or Twitter is not a good indication of an actual friendship relationship between the individuals. So, to identify who's friends with whom, we need to look at the interaction between individuals. Our classification identifies one-on-one interactions, which are more informative than interactions in the context of a group.

10. REFERENCES

[1] R. Albert and A.-L. Barabasi. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[2] L. Backstrom, R. Kumar, C. Marlow, J. Novak, and A. Tomkins. Preferential behavior in online groups. In *Proc. 1st WSDM*, pages 117–128, 2008.

[3] L. Barabiási. The origin of bursts and heavy tails in human dynamics. *Nature*, 435, 2005.

[4] S. Bikhchandani, D. Hirshleifer, and I. Welch. A theory of fads, fashion, custom, and cultural change in informational cascades. *Journal of Political Economy*, 100(5):992–1026, 1992.

[5] B. Bollobas. *Random Graphs*. Cambridge, 2001.

[6] B. Bollobas and O. Riordan. *Mathematical Results on Scale-Free Random Graphs*, pages 1–37. Wiley–WCH, 2002.

[7] H.-C. Chen, M. Magdon-Ismail, M. Goldberg, and W. A. Wallace. Inferring agent dynamics from social communication network. In *Proc. 9th WebKDD*, 2007.

[8] Z. Dezsö, E. Almaas, A. Lukács, B. Rácz, I. Szakadát, and A.-L. Barabási. Dynamics of information access on the web. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 73(6):066132, 2006.

[9] R. Durrett. *Random Graph Dynamics*. Cambridge, 2006.

[10] D. Fisher, M. Smith, and H. T. Welser. You are who you talk to: Detecting roles in usenet newsgroups. In *Proc. 39th HICSS*, 2006.

[11] J. Goldenberg, B. Libai, and E. Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, 12(3):209–221, 2001.

[12] B. Golub and M. O. Jackson. The power of selection bias in explaining the structure of observed Internet diffusions. *Proc. National Academy of Sciences*, To appear.

[13] M. Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, 83(6):1420–1443, 1978.

[14] T. E. Harris. *The Theory of Branching Processes*. Dover, 2002.

[15] H. S. Heaps. *Information Retrieval: Computational and Theoretical Aspects*. Academic Press, 1978.

[16] J. Kleinberg. Complex networks and decentralized search algorithms. In *Proc. International Congress of Mathematicians*, 2006.

[17] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *Proc. 12th KDD*, pages 611–617, 2006.

[18] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. In *Proc. 41st FOCS*, pages 57–65, 2000.

[19] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. In *Proc. 7th EC*, pages 228–237, 2006.

[20] J. Leskovec, L. Backstrom, and J. M. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proc. 15th KDD*, pages 497–506, 2009.

[21] J. Leskovec, L. Backstrom, R. Kumar, and A. w. Tomkins. Microscopic evolution of social networks. In *Proc.14th KDD*, pages 462–470, 2008.

[22] J. Leskovec, J. M. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *TKDD*, 1(1), 2007.

[23] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs. In *Proc. 7th SDM*, 2007.

[24] J. Leskovec, A. Singh, and J. Kleinberg. Patterns of influence in a recommendation network. In *Proc. 10th PAKDD*, pages 380–389, 2006.

[25] D. Liben-Nowell and J. Kleinberg. Tracing the flow of information on a global scale using Internet chain-letter data. *Proc. National Academy of Sciences*, 105(12):4633–4638, 2008.

[26] M. McGlohon and M. Hurst. Community structure and information flow in usenet: Improving analysis with a thread ownership model. In *Proc. 3rd ICWSM*, 2009.

[27] M. Mcglohon, J. Leskovec, C. Faloutsos, M. Hurst, and N. Glance. Finding patterns in blog shapes and blog evolution. In *Proc. 1st ICWSM*, 2007.

[28] M. Molloy and B. Reed. A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms*, 6(2/3):161–180, 1995.

[29] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge, 1995.

[30] T. C. Turner, M. A. Smith, D. Fisher, and H. T. Welser. Picturing usenet: Mapping computer-mediated collective action. *Journal of Computer-Mediated Communication*, 10(4), 2005.

[31] A. Vazquez. Spreading dynamics on heterogeneous populations: multi-type network approach. *Physical Review Letters*, 74, 2006.

[32] F. B. Viegas and M. Smith. Newsgroup crowds and authorlines: Visualizing the activity of individuals in conversational cyberspaces. In *Proc. 37th HICSS*, page 10, 2004.

[33] D. J. Watts. A simple model of global cascades on random networks. *Proc. National Academy of Sciences*, 99(9):5766–5771, 2002.