Global Dynamics of Online Group Conversations

Rushi Bhatt
Yahoo! Labs
Bangalore, India 560071
rushi@yahoo-inc.com

Kishor Barman
Tata Institute of Fundamental Research
Mumbai, India 400 005
kishor@tcs.tifr.res.in

Abstract

Public online groups allow individuals to carry out conversations of common interests. Study of such group conversations provides a unique opportunity to study patterns of human conversations without violating individual privacy. The observational studies conducted in this paper are an attempt to identify the main correlates of continued growth of conversations, thereby clearing the path to developing predictive models user participation.

We study temporal evolution of online group discussions. Surprisingly, we find that individual discussion groups display distinctively q-exponential shaped inter-message times to reply distributions, unlike the power law distributions seen in email conversations. We show, using simulations, that the heavy-tailed distribution of time to reply, which we also observe when all data is combined, originate from mixtures of q-exponentials. We also find that popular threads come to be so from the very beginning as opposed to evolving to be more popular as they grow. This raises new possibilities for developing generative models of thread growth.

Introduction

Online groups are an important channel of social interaction that facilitate topic-specific discussions among a limited set of individuals. Online groups thus fill the gap between person-to-person social channels like email and Instant Messages (IMs), and broadcast channels like Twitter and Facebook. In this paper, we take a global account of how group message threads evolve with time as a function of the social identities of the participants.

Our contributions:

We find that good threads in fact start off well, rather than evolving to be so! This is observation is not natural to the usual preferential attachment models which attach higher probabilities of replies to threads that have already become large.

We also analyze temporal characteristics of online group threads using a large data set. We find that the time to reply distributions have strong circadian modulations. We propose that the q-exponential distribution is a better parametric fit for individual groups, and show that a mix of individual q-exponentials gives rise to the familiar power-law like time to reply distributions.

Related Work

Structure and growth of conversations:

(Backstrom et al. 2008) showed that highly engaged users in online groups start out differently from others: Users who go on to be engaged and highly active are received positively by other groups users *from the beginning*. Similar to the findings of Backstrom et al., we observe that highly successful conversation threads also start out differently from others and are mostly dependent more on identity of the originator of threads and the initial rate at which they receive messages rather than the users who post replies to original messages. In hindsight, this may be obvious, since the initial rate at which replies are posted may be a correlate of the topic or the social status of the thread originator.

Inter-message time analysis:

Power-law distributions of time-delays between online social activities have been widely observed. On a large database of email conversations, Johansen (Johansen 2004) empirically showed a power-law fit for the time-difference between successive email replies. Barabasi (Barabasi 2005) proposed a priority-queue based model to explain the power-law fit observed in those email conversations. Malmgren and colleagues (Malmgren et al. 2008) showed that it is also plausible that heavy tailed reply time distributions arise as a mixture of inhomogeneous Poisson distributed arrival times. They model user posting rates as inhomogeneous Poisson distributions where the rate parameter is modulated by circadian rhythms. They show that a heavy tailed distribution of reply times arises when many individual inhomogeneous Poisson distributions are mixed. While we do not model individual users here, we draw inspiration from the work of Briggs and Beck (Briggs and Beck 2006) where the authors model train delays as q-exponentials and show how heavy-tailed distributions (e.g, q-exponentials) arise naturally as superstatistics of individual exponential (e.g., Poisson) processes with randomly distributed (e.g., χ^2) rate parameters.

Lescovec et al. (Leskovec et al. 2007) showed how an approximate power-law distribution was also present in the delay between blog posts linking to each-other, and provided a generative law that gave rise to the observed power law distributions.

Preliminaries

Data description

We use one of the world's largest collections of online discussion boards. From this Groups data we chose only the public groups that were moderated and active at the time of data collection. Of these, we retained groups with at least 10 total messages and at least 10 unique subscribers. Roughly 10,000 groups with about 14 million total messages remained after filtering. The data was collected in January 2010 and contained all messages in the selected groups up to that time, spanning nearly 10 years.

While individual users may receive updates through different channels (email, consolidated digests, direct website visits) we do not discriminate between these in our analysis.

Temporal dynamics of message posting

In this section, we will present an overview of some temporal characteristics of time to reply differences that seem unique to discussion groups.

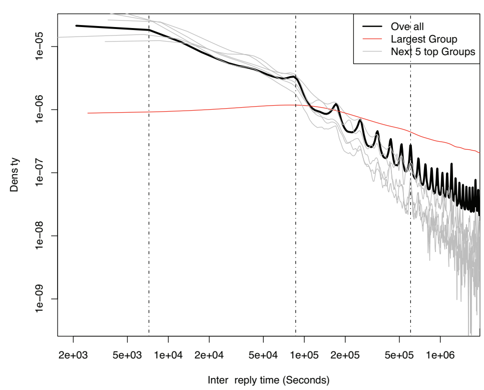
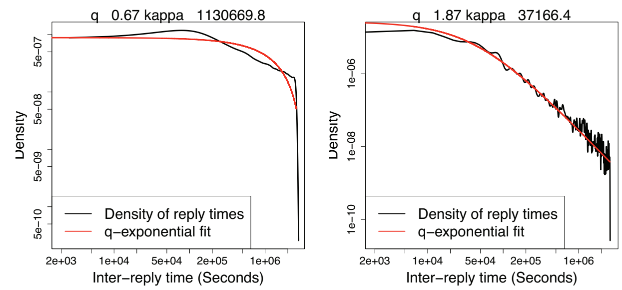


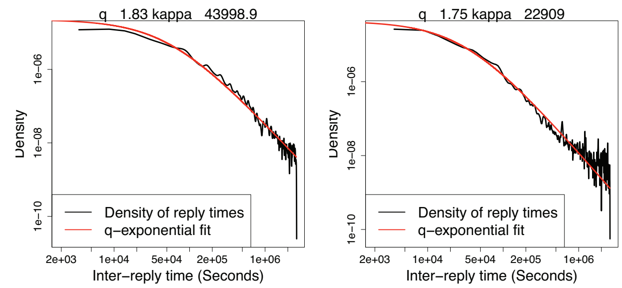
Figure 1: Time delay between successive messages is distinctly non-power law. We also see a strong circadian rhythm and "knee points" at two hours, and one day, followed by successive daily peaks (dashed vertical lines at 2 hrs, 1 day, 1 week)

q-exponentials describe time to reply distributions

The familiar power law distribution of the size of threads exists, with probability of observing a thread of size k following distribution $p(k) \propto k^{-\alpha}$ for some α. On the other hand, the distribution of times to reply $r(v, u)$ for individual groups, and to an extent over the whole data, follows a distinctively non-power law distribution. For example, Figure 1 shows the top 6 individual groups show lighter tailed distributions (gray lines) than what a power law distribution would predict. However, time to reply distribution accumulated over all the groups may still be closer to a power law (dark line in Figure 1). We also observe strong circadian



(a) Group 1 (b) Group 2



(c) Group 3 (d) Group 4

Figure 2: Density of the time to reply distributions of the individual groups, and their Maximum Likelihood fit to q-exponentials.

modulations in time to reply distributions. In the overall time to reply distribution across all groups we observe distinct knee points at two hours and one day (See Figure 1). Keeping the distribution between two hours and one day as a reference, we see that there is lower mass in the distribution at less than two hours and more than one day compared to what a power law would predict. We observe more prominent circadian modulation than in some email data sets (Malmgren et al. 2008) and see traces of a weekly modulation as well, similar to blog link creation time course (Leskovec et al. 2007).

Visually, the time to reply distributions for individual groups seem to be q-exponentials. A random variable X is q-exponentially distributed with shape and scale parameters q and κ, respectively, if its upper cumulative (or complementary) distribution function is $Pr[X \geq x] = \left(1 - \frac{(1-q)x}{\kappa}\right)^{\frac{1}{1-q}}$ (Picoli, Mendes, and Malacarne 2003). The q-exponential distribution has been used in the literature for modeling social network growth (White et al. 2006).

In Figure 2 we show reply time distributions for some of the largest groups and their corresponding q-exponential fits obtained by a Maximum Likelihood estimate (Shalizi 2007).

Mixture of q-exponentials result in power law From the above observations, we postulate that for individual groups, time to reply are q-exponential distributed. We also postulate that when a right mix of q-exponentials is accumulated over all individual groups, it is possible to generate an overall dis-

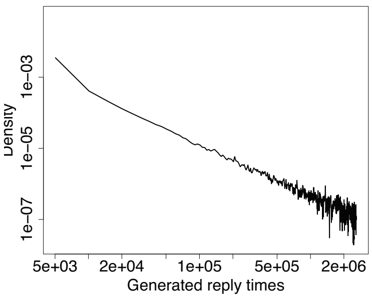
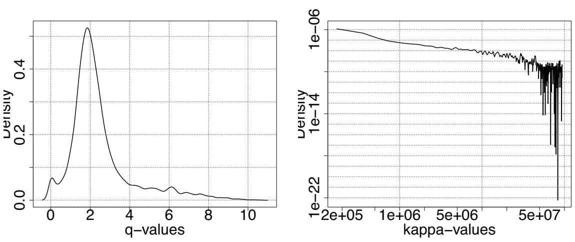
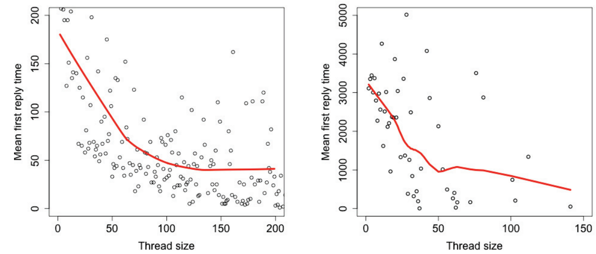


Figure 3: Power law distribution of simulated time to reply aggregated over all the groups (1000 groups), when the time to reply for each groups is generated using q-exponentials.

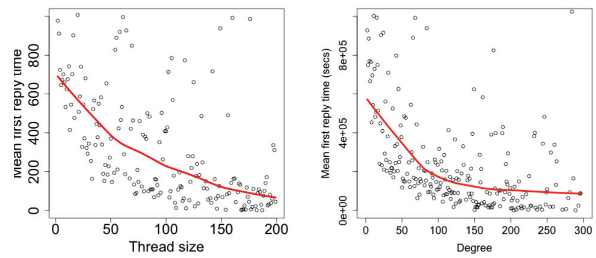


(a) q-values (b) κ-values

Figure 4: Distribution of the parameters of q-exponential fit across groups.



(a) Group 1 (b) Group 2



(c) Over all groups (d) Mean first reply time vs. Degree

Figure 5: (a,b,c) Thread size vs mean first reply time for two largest groups and over all the groups. The X axis represents thread size, and the Y axis represents mean first reply times. (b) Messages that receive many replies also get a quicker first reply. The red curves indicate LOWESS smoothening.

tribution close to a power-law. For example, Figure 3 shows a resulting distribution when we mix 1000 independent q-exponential distributions. For our simulations, we first estimated individual group-level q-exponential parameters. We then sampled 1000 points from these parameter distributions and sampled equal number of samples from the distributions governed by these parameters. When samples across all groups are merged, these individual q-exponential distributions give rise to a power law like distribution (see Figure 3).

Figure 4 shows the distributions of shape parameter q (Figure 4(a)) and the scale parameter κ (Figure 4(b)) estimated by Maximum Likelihood estimates for each of the 1000 groups.

Temporal Structure of Threads

We now turn to understanding the characteristics of individual threads. We would like to find out what may be the drivers of user participation, and whether there are any social attributes of ongoing threads that attract increased and quicker participation.

Popular threads generate high interest from start A popular thread could be popular from the start, or it could start slow and gain popularity at a later stage. We want to understand which of these possibilities prevail. To that end we look at how time to *first reply* to the root message relates to the eventual thread size. Figure 5 shows the average time to first reply to a root message vs. thread size. Generally, first replies to the root message arrive much quicker for threads that grow to receive many replies. This is true within a group (not shown here due to space constraints), as well as aggregated over all the groups (Figure 5(c)). Kumar et al. (Kumar, Mahdian, and McGlohon 2010) also observed that the exponent of degree distribution power law decreases with depth in threads, i.e., thread trees tend to have fewer answers at higher depths. This suggests that popular threads usually are popular from the start, and begin receiving quicker replies right from the time the root message is posted. In other words, the root message content or the identity of its author determine, to a large extent, the eventual success of a thread.

First reply time predicts degree We would also expect that if a message receives a quick first reply, then probably it is interesting enough to receive many more subsequent replies. Indeed this is true. Figure 5(d) shows how messages with higher degree also generally have quicker first reply times.

Mean reply time over the whole thread has no systematic relation to thread size We saw how time to *first reply* to the root message correlates well with the eventual thread size. The *average delay* over all replies to a message, on the other hand, paints a completely different picture. For many groups, we in fact see an increase in the mean reply time

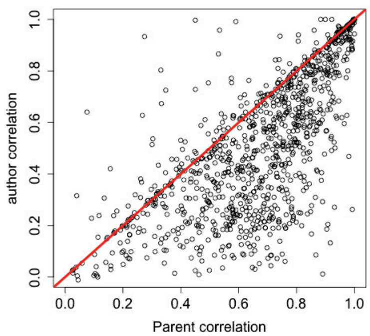


Figure 6: Across groups, baseline reply-receive times of parent-node authors are better correlates of time-to-reply

with thread size (not shown here due to space constraints).

We also see that for some groups, the mean reply time oscillates unpredictably as the thread size grows. We looked at a large number of threads but found no systematic pattern in thread size vs. mean times to reply.

Parent author is more predictive of time to reply For each group G as we compute two correlation coefficients. First, we correlate the time to reply with the baseline posting rates of authors of *replies*, called the *author baseline*. Specifically, for each message pair $(parent(v), v)$ posted in group G we compute Pearson's correlation coefficient $R_a(G)$ between quantities $r(parent(v), v)$ and $author\_baseline(author(v))$. Second, we correlate the time to reply with the average rate at which a given user's posts *receive* replies, called the *parent baseline*. In other words, we compute Pearson's correlation coefficient $R_p(G)$ between $r(parent(v), v)$ and $parent\_baseline(parent\_author(v))$. As seen in Figure 6, the *parent baseline* correlates better with time-to-replies in a thread than the *author baseline*

Discussion and Conclusions

The evolution of conversations when limited amount of information is visible to users is not yet completely characterized. Most of the generative models in the literature explicitly use the current degree (the number of friends a user has, total replies so far) for creating new edges. It is unknown whether users in fact take this information explicitly into consideration, or always have access to this information while creating links. Preferential attachment models are thus at best a proxy when current degree is not available to users during link creation. These aspects of evolutions of conversations remain to be investigated. In this paper, we showed how the initial few posts in a thread matter in the ultimate size of threads, and how there are non-trivial temporal aspects of thread evolution that are not captured by popular models.

References

Backstrom, L.; Kumar, R.; Marlow, C.; Novak, J.; and Tomkins, A. 2008. Preferential behavior in online groups. In *Proceedings of the international conference on Web search and web data mining*, WSDM '08, 117–128. New York, NY, USA: ACM.

Barabasi, A.-L. 2005. The origin of bursts and heavy tails in human dynamics. *NATURE* 435:207.

Briggs, K., and Beck, C. 2006. Modelling train delays with q-exponential functions. *arXiv*.

Johansen, A. 2004. Probing human response times. *Physica A: Statistical Mechanics and its Applications* 338(1-2):286 – 291. Proceedings of the conference A Nonlinear World: the Real World, 2nd International Conference on Frontier Science.

Kumar, R.; Mahdian, M.; and McGlohon, M. 2010. Dynamics of conversations. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, 553–562. New York, NY, USA: ACM.

Leskovec, J.; McGlohon, M.; Faloutsos, C.; Glance, N. S.; and Hurst, M. 2007. Patterns of cascading behavior in large blog graphs. In *SDM*. SIAM.

Malmgren, R. D.; Stouffer, D. B.; Motter, A. E.; and Amaral, L. A. N. 2008. A Poissonian explanation for heavy tails in e-mail communication. *Proc. Natl. Acad. Sci. U.S.A.* 105:18153–18158.

Picoli, S.; Mendes, R. S.; and Malacarne, L. C. 2003. q-exponential, weibull, and q-weibull distributions: an empirical analysis. *Physica A: Statistical Mechanics and its Applications* 324(3-4):678 – 688.

Shalizi, C. R. 2007. Maximum Likelihood Estimation for q-Exponential (Tsallis) Distributions. *arXiv* math/0701854v2.

White, D. R.; Kejžar, N. c. v.; Tsallis, C.; Farmer, D.; and White, S. 2006. Generative model for feedback networks. *Phys. Rev. E* 73(1):016119.