Diversity of User Activity and Content Quality in Online Communities

Tad Hogg and **Gabor Szabo**
HP Labs
Palo Alto, CA

Abstract

Web sites where users create and rate content display long-tailed distributions in many aspects of behavior. Using one such community site, Essembly, we propose and evaluate mechanisms to explain these behaviors. Unlike purely descriptive models, these mechanisms rely on user behaviors based on information available to each user. For Essembly, we find the long-tails arise from large differences among user activity rates, the time users devote to the site, and qualities of the rated content. The models not only explain overall behavior but also allow estimating the properties of users and content from their early behaviors.

Introduction

Participatory web sites facilitate their users creating, rating and sharing content. Examples include Digg[.com] for news stories and Wikipedia[.org] for encyclopedia articles. To aid users in finding content of interest to them, many such sites employ collaborative filtering of the content (Lam 2004) to allow users to specify links to other users whose content or ratings are particularly relevant. These links can involve either people who already know each other (e.g., friends) or people who discover their common interests through participating in the web site. The resulting networks enable users to find others with similar interests and establish trust in recommendations (Guha et al. 2004).

The availability of activity records from these sites has led to numerous studies of user behavior and the networks they create. Observed commonalities in these systems have identified general generative processes leading to these observations. Examples include preferential attachment in forming networks and multiplicative processes in rating content, leading to wide variation in behaviors. While such models provide a broad understanding of the observations, they often lack causal connection with individual user behaviors based on user preferences and the information available to users in making their decisions (Vázquez 2003; Boccaletti et al. 2006). Moreover, observed behavior can arise from a variety of mechanisms (Mitzenmacher 2004).

Predicting consequences of alternate designs of the web site requires models of causal behavior. Establishing such models is difficult due to the possibility of unobserved confounding causal factors. Instead, such models would ideally be based on intervention studies and randomized trials to identify important causal relationships. In contrast to the wide availability of observational data on user behavior, such intervention studies are difficult, though this is situation is improving with the increasing feasibility of experiments in large virtual communities (Bainbridge 2007) and large-scale web-based experiments (Salganik, Dodds, and Watts 2006).

A simpler, though less conclusive, approach to causal models is limiting the mechanisms to use information readily available to users on a participatory web site. Such models provide specific hypotheses to test with future intervention experiments and also suggest improvements to the web site by altering the user experience, e.g., available information or incentives. The simplest such approach considers average behavior of users on a site (Lerman 2007a). Such models relate system behavior to the average decisions of many users. By design, such models do not address a prominent aspect of observed online networks: the long tails in their distributions. Models including this diversity could help improve effectiveness of the web sites by allowing focus on significantly active users or especially interesting content, and enhancing user experience by leveraging the long tail in niche demand (Anderson 2006).

A key question with respect to the observed diversity is whether user behavior and content characteristics are reasonably viewed as arising from a statistically homogeneous population, and hence well-characterized by a mean and variance. Or is diversity of prior intrinsic characteristics among participants the dominant cause of the observed wide variation in behaviors? In the latter case, can these characteristics be estimated (quickly) from (a few) observations of behavior, allowing site designers to use estimates of these characteristics, e.g., to highlight especially interesting content? Moreover, to the extent user diversity is important, what characterization of this user variation is sufficient to produce the observed long-tail distributions?

This paper considers these questions in the context of a politically-oriented web community, Essembly[1]. We consider population diversity and mechanisms users could be following to produce the observed long-tail behaviors. In the

[1] Essembly LLC at www.essembly.com

remainder of this paper, we first describe Essembly and our data set. We then examine highly variable behaviors as well as models for users and content rating. With these models in hand, we consider their possible use during operation of the web site by helping identify user and content characteristics early in their history. Finally we discuss implications and extensions to other participatory web sites.

Essembly

Essembly is an online service helping users engage in political discussion through creating and voting on *resolves* reflecting controversial issues. Essembly provides three distinct networks for users: a social network, an ideological preference network, and an anti-preference network, called *friends*, *allies* and *nemeses*, respectively.

The distinct social and ideological networks enable users to distinguish between people they know personally and people encountered on the site with whom they tend to agree or disagree. The Essembly user interface presents several options for users to discover new resolves, e.g., based on votes by network neighbors, recency, overall popularity, and degree of controversy.

Our data set consists of anonymized voting records for Essembly between its inception in August 2005 and December 2006, and the users and links in the three networks at the end of this period. Our data set has $15,424$ users. Essembly presents 10 resolves during the user registration process to establish an initial ideological profile used to facilitate users finding others with similar or different political views. To focus on user-created content, we consider the remaining $24,953$ resolves, with a total of 1.3 million votes. The structure of the networks is typical of those seen in online social networking sites, and the links created by users generally conform to their nominal semantics (Hogg et al. 2008).

In common with other sites, such as Digg and Youtube, the Essembly data shows wide ranges in user participation, interest in content (the resolves) and degree distribution in networks. Essembly is a relatively small site, for which it is feasible to examine the behavior of all users and all content during our sample period.

Users

A key quantity characterizing the users is the length of time they choose to remain active on the site. We measure activity time as the time between a user's first and last votes (this includes votes on the initial resolves during registration – users need not vote on all of them immediately). Most users are active for only a short time (less than a day). The 4762 users active for at least a day account for most of the votes and links, and we focus on these *active users* for our model.

Fig. 1 shows the distribution of the activity times of active users. Specifically, a point at time t in the figure is the fraction of active users who remained active at least $t + 1$ days from among active users who joined Essembly at least $t + 1$ days before the end of our sample period. Only this set of early users could have remained active $t + 1$ or more days within our sample. The points at larger times have larger error bars because they involve smaller sets of users who joined the site early enough to include in the estimate. The figure shows a Weibull distribution fit to the activity times. This indicates a mixture of processes (Frisch and Sornette 1997) leading users to abandon the site: the longer a user remains active, the lower the probability per unit time they become inactive, as also occurs in a variety of other web sites (Wilkinson 2008).

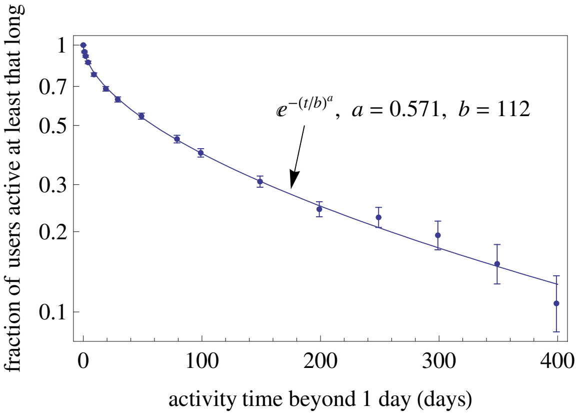

Figure 1: Distribution of activity times for active users on a log plot. The line shows a Weibull distribution fit to the values, equal to $e^{-(t/b)^a}$ with parameters in Table 1. Error bars show the 95% confidence intervals.

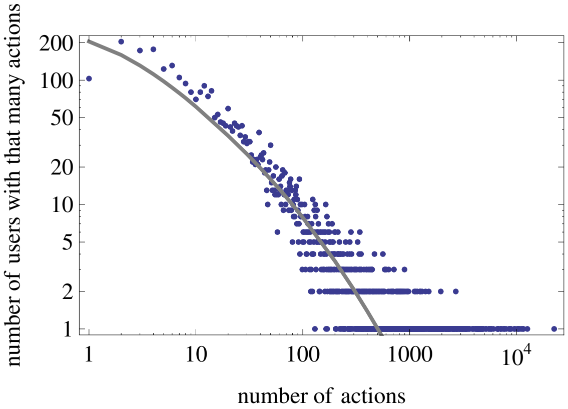

Figure 2: Distribution of number of active users vs. the number of actions (votes and links) a user made. The solid curve is the distribution from the model described in the text. The plot does not include the 21 users with no votes or links.

User actions consist of *voting*, *creating resolves*, and *forming links*. Fig. 2 shows the distribution of number of actions among the users. This distribution arises from two factors: how long users participate on the site, and how often they act on the site while active. These properties are only weakly correlated (correlation coefficient -0.07 among active users).

Model

Fig. 3 summarizes our model for users' participation and their activities on the site while they are active. New users

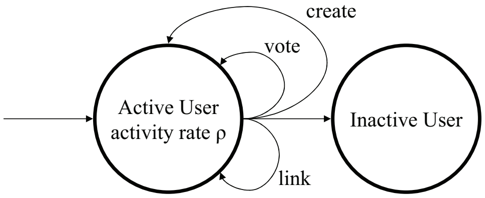

Figure 3: Model of user behavior. People join the site as active users, who *create* resolves, *vote* on them and *link* to other active users. Users can eventually stop participating and become inactive.

parameter	value
activity time constants	$a = 0.571 \pm 0.046$
	$b = 112 \pm 13$ days
resolve creation	$q = 0.018 \pm 0.0002$
link creation	$\lambda = 0.043 \pm 0.0003$

Table 1: User activity parameters. The ranges specify 95% confidence intervals for the parameter estimates.

arrive in the system when they register. Users leave the system (*i.e.*, become inactive) with a rate decreasing with the time they have been active. Specifically, as seen in Fig. 1, a user active for at least a day has total activity time t beyond that one day given by a Weibull distribution, i.e., $P_{\text{time}}(t) = ab^{-a}t^{a-1}e^{-(t/b)^a}$. Table 1 gives the values for these model parameters.

User activity is clumped in time, with groups of many votes close in time separated by gaps of at least several hours. This temporal structure can be viewed as a sequence of user sessions. The averaged distributions for interevent times between activities of individuals show long-tail behavior, similar to other observed human activity patterns, such as email communications or web site visits (Vázquez et al. 2006). To model the numbers of actions per user in the long time limit where we are only interested in the total number of accumulated votes for a particular user, this clumping of actions in time is not important. Specifically we suppose each user has an average activity rate ρ while they are active on the site, whose maximum likelihood estimate is $\rho_u = e_u/T_u$, where ρ_u is user u's activity, e_u is that user's number of events (i.e., votes, resolve creations and links), and T_u is the time elapsed between the user's first and last vote. We suppose the ρ_u values arise as independent choices from a distribution $P_{\text{user}}(\rho_u)$ and the values are independent of the length of time a user is active on the site, corresponding to the near zero correlation between activity time and votes mentioned above.

Voting is by far the most common user activity. We characterize these individual activities by fractions q and λ for creating resolves and forming links, respectively. The rate of voting on existing resolves for a user is then $\rho_u(1-q-\lambda)$. The observed values of q and λ vary somewhat among users. With our focus on a user's total activity on the site, we report average values for q and λ and examine the variation among users due to their differing overall activity rates ρ_u and amount of time they are active on the site T_u.

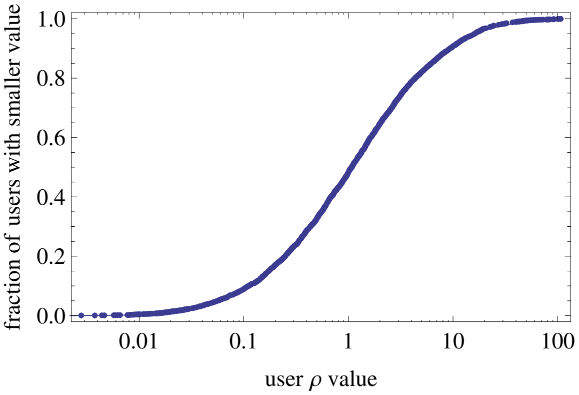

Figure 4: Cumulative distribution of ρ_u values for the 4741 users who were active at least one day and voted on at least one noninitial resolve or formed at least one link. Plot includes a curve for a lognormal distribution fit, which is indistinguishable from the points and with parameters $\mu = 0.03 \pm 0.05$ and $\sigma = 1.70 \pm 0.04$. The ρ values are in units of actions per day. In this and other figures the range given with the parameter estimate is the 95% confidence interval.

Behavior

We estimate the model parameters from the observed user activities, and restrict attention to active users. Table 1 shows the estimates for parameters, q and λ, governing activity choices. Fig. 4 shows the observed cumulative distribution ρ_u values and a fit to a lognormal distribution. The parameter estimates and confidence intervals in this and the other figures are maximum likelihood estimates (Newman 2005; James and Plank 2007).

In this model, the probability user u has e_u events (votes, resolve creations, and links) is a Poisson distribution with mean $\rho_u T_u$ where T_u is the time the user is active. For users joining the system near the end of our sample, T_u is limited by the end of the sample rather than when the user decides to become inactive.

The heavy-tailed nature of the actions per user distribution (Fig. 2) can be attributed to the interplay between the user activity times T_u and the broad lognormal distribution of the user activity rates ρ_u: the product of these two distributions (Glen, Leemis, and Drew 2004) results in a broad distribution of the product $\rho_u T_u$. In particular, the lognormal distribution fit to the ρ values and the activity times of the users give a distribution of means for Poisson distributions of number of actions for the users. This combination gives the distribution shown as the curve in Fig. 2. Thus this model accounts for the extended tail of the activity. For users with relatively little activity, the model underestimates the number of users with about 2 to 20 actions while overestimating the number with with zero or one action. This arises from a dependence between activity time and activity rate for these less active users. In particular, such users have negative correlation coefficient between their activity rate and time, tending to increase their number of actions.

The distributions of activity times and rates reflect the

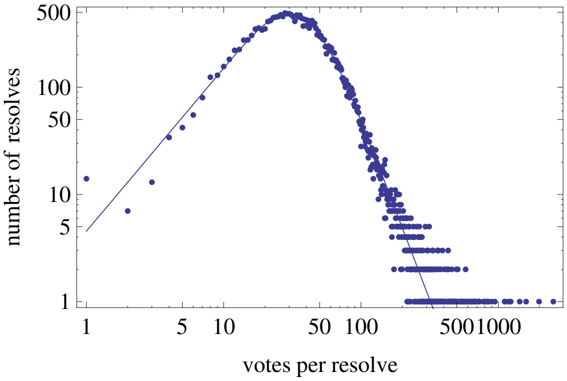

Figure 5: Distribution of votes on resolves. The solid curve indicates a double Pareto lognormal fit to the values, with parameters $\alpha = 2.4 \pm 0.1$, $\beta = 2.5 \pm 0.1$, $\mu = 3.67 \pm 0.02$ and $\sigma = 0.38 \pm 0.02$.

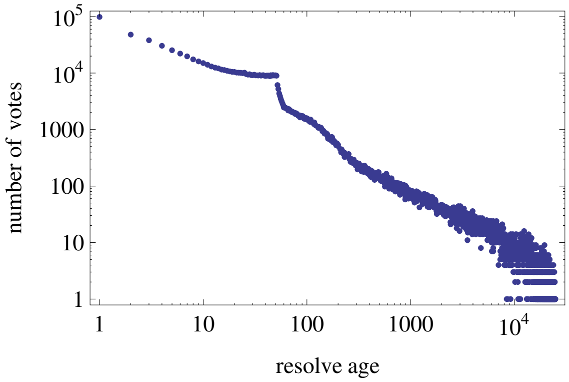

Figure 6: Distribution of votes vs. age of a resolve.

range of dedication of users to the site. Most users try the service for a only a short time while active users give the heavy tail. Such extended distributions of user activity rates are also seen in other web sites, such as Digg (Lerman 2007b; Wilkinson 2008), and in activities such as scientific productivity (Shockley 1957).

Resolves

A key question for user-created content is how user activities distribute among the available content. For Essembly, Fig. 5 shows the broad distribution in total number of votes per resolve. In Essembly, each resolve receives its first vote when it is created, i.e., the vote of the user introducing the resolve. Thus the observed votes on a resolve are a combination of two user activities: creating a new resolve (giving the resolve its first vote) and subsequently other users choosing to vote on the resolve if they see it while visiting the site.

We consider a user's selection of an existing resolve to vote on as mainly due to a combination of two factors: visibility and interestingness of a resolve to a user. Visibility is the probability a user finds the resolve during a visit to the site. Interestingness is the conditional probability a user votes on the resolve given it is visible to that user. These two factors apply to a variety of web sites, e.g., providing a description of average behavior on Digg (Lerman 2007a).

The web site's user interface design determines content visibility. Typically sites, including Essembly, emphasize recently created content and popular content (i.e., receiving many votes over a period of time). Essembly also emphasizes controversial resolves. As with other networking sites, the user interface highlights resolves with these properties both globally and among the user's network neighbors. Users can also find resolves through a search interface.

For Essembly, the networks have only a modest influence on voting (Hogg et al. 2008). Recency appears to be the most significant factor affecting visibility. Fig. 6 shows how votes distribute according to the age of the resolve at the time of the vote. We define the *age* of the resolve as the ordinality of the given resolve among resolves introduced in time. An age 1 resolve is the newest one of the resolves introduced, while the oldest resolve has age R where R is the number of resolves. Most votes go to recent resolves, i.e., those with a small age.

The decay in votes with age is motivated by recency (decreasing visibility with age as resolve moves down, and eventually off, the list of recent resolves). We offer no underlying model for this "aging function" but its overall power-law form corresponds to users' willingness to visit successive pages or scroll down a long list (Huberman et al. 1998). The step at age 50 is, presumably, due to a limit on number of recent resolves readily accessible to users. The values decrease as a power law, proportional to a^{-s}, where a is resolve age and $s \approx 0.5$ up to about age 50. For larger ages, the values in Fig. 6 decreases faster, with $s \approx -1$.

The combination of different ages in the data sample is a significant factor in producing the observed distribution of votes (Huberman and Adamic 1999). In particular, a distribution of ages and a multiplicative process produces a lognormal distribution with power-law tails, the double Pareto lognormal distribution (Reed and Jorgensen 2004), with four parameters. Two parameters, μ and σ characterize the location and width of the center of the distribution. The remaining parameters characterize the tails: α for the power-law decay in the upper tail, with number of resolves with v votes proportional to $v^{-\alpha-1}$, and β for the power-law growth in the lower tail, with number of resolves proportional to $v^{\beta-1}$. Fig. 5 shows a fit of this distribution to the numbers of votes different resolves received.

Model

Our model of resolve creation involves a fraction q of each user's activity on the site, on average, giving each resolve its first vote. For subsequent votes, we view a user's choice of resolve as due to an intrinsic "interestingness" property r of each resolve and its visibility.

In general, r could depend on the resolve age and its popularity (especially among network neighbors, if neighbors influence a user to vote rather than just make a resolve more visible). However, for simplicity, we take r to be constant

	interval			
resolve	I_1	I_2	I_3	I_4
1	$f(1)r_1v_1$	$f(2)r_1v_2$	$f(3)r_1v_3$	$f(4)r_1v_4$
2	–	$f(1)r_2v_2$	$f(2)r_2v_3$	$f(3)r_2v_4$
3	–	–	$f(1)r_3v_3$	$f(2)r_3v_4$
4	–	–	–	$f(1)r_4v_4$

Table 2: Model of distribution of votes among resolves in time intervals between successive resolve introductions, here shown for the first four resolves.

for a resolve. A key motivation for this choice is the observation that high or low rates of voting on a resolve tend to persist over time, when controlling for the age and number of votes the resolve already has. Thus a constant value for an intrinsic interestingness property of resolves is a reasonable approximation for Essembly. Since we use these r values to model behavior of users other than the person who created the resolve, there is no need to consider separately the high interest in the resolve presumably reflected by the creator's choice to introduce the resolve. Thus we further assume r is independent of the user, which amounts to considering general interest in resolves among the population rather than considering possible niche interests among subgroups of users. With these simplifications, we take the r values to arise as independent choices from a distribution $P_{\text{resolve}}(r)$.

Visibility of a resolve depends on age, rank in number of votes compared with other resolves (popularity), controversy, both in general and among user's neighbors. For Essembly, resolve age appears to be the most significant factor, so we take visibility to be a function of age alone, i.e., determined by a function $f(a)$.

With these factors, we model the chance that the next vote on existing resolves goes to resolve x as being proportional to $r_x f(a_x)$ where a_x is the age of the resolve at the time of the vote. The model's behavior is unchanged by an overall multiplicative constant, and we arbitrarily set $f(1) = 1$.

Behavior

Our model of resolve votes requires estimating the distribution P_{resolve} and the aging function $f(a)$. To do so, we consider the votes (other than the first vote on each resolve) between successive resolve introductions. Let R be the number of resolves in our data sample. We denote the resolves in the order they were introduced, ranging from 1 to R.

Let v_i be the number of votes made in the time interval I_i between the introductions of resolves i and $i+1$ (not including the two votes accompanying those resolve introductions). During this interval, the system has i existing resolves. When the number of existing resolves is large, we can treat the votes going to each resolve as approximately independent. In this case, the number of votes resolve $j \leq i$ receives during time interval I_i is approximately a Poisson process with mean $v_i r_j f(i-j+1)$ because during this interval resolve j is of age $i-j+1$. Table 2 illustrates these relationships.

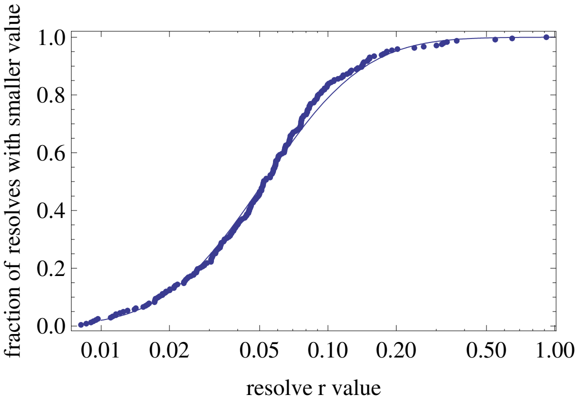

Figure 7: Cumulative distribution of r values for a sample of resolves in the middle of our data set. The curve shows a lognormal distribution fit, with parameters $\mu = -3.0 \pm 0.1$ and $\sigma = 0.80 \pm 0.07$.

We estimate the r_i and $f(a)$ values as those maximizing the likelihood of the observed numbers of votes on the resolves in these time intervals, treated as independent Poisson processes. Since we are interested in estimating the distribution $P_{\text{resolve}}(r)$ rather than the r_i values of all the resolves, it is sufficient to maximize over a sample of resolves and time intervals from the middle of our data set (when there are thousands of existing resolves and votes cast). The resulting $f(a)$ estimate is similar to the distribution of votes vs. age in Fig. 6, and Fig. 7 shows the distribution of estimated r values and a lognormal fit.

With the wide variation in r values and the activity rates for users (Fig. 4), a natural question is whether these variations are related. In particular, whether the most active users tend to preferentially introduce resolves that are especially interesting to other users. While active users tend to introduce more resolves overall, the correlation between the activity rate of a user and the average r values of the resolves introduced by that user is small: -0.06. We find a modest correlation (0.16) between the *time* a user is active on the site and the mean r values of that user's introduced resolves.

To relate this model to the vote distribution of Fig. 5, consider the votes received by resolve j up to and including the time it is of age A. According to our model, the number of votes, *other than* its first vote, this resolve receives is a Poisson variable $V_j(A)$ with mean

$$\mu_j(A) = r_j \sum_{a=1}^{A} f(a) v_{j+a-1}$$

At the end of our data set, resolve j is of age $R - j + 1$.

The persistence of votes on resolves based on the wide variation of r values among resolves gives rise to a multiplicative process with decay. To see this, in our model the number of votes between successive resolve introductions is geometrically distributed with mean $\hat{v} = (1-\lambda)/q - 1 \approx 50$. Furthermore, from Fig. 6, the aging function is approximately power law, with $f(a) \approx a^{-s}$ and $\sum_{a=1}^{A} f(a) \sim A^{1-s}/(1-s)$. The expected number of votes up to age

A is then $\mu_j(A) \sim r_j \hat{v} A^{1-s}/(1-s)$. After accumulating many votes (i.e., when A is large), the actual number of votes $V_j(A)$ will usually be close to this expected value. The change in votes to age $A+1$ is

$$\begin{aligned} V_j(A+1) &\approx r_j \hat{v} \frac{A^{1-s}}{1-s}(1+x) \\ &\approx V_j(A)(1+x) \end{aligned}$$

where x is a nonnegative random variable with mean $(1-s)/A$. Thus, except possibly for the votes a resolve receives shortly after its introduction, the growth in number of votes is well-described by a multiplicative process with decay.

That our model corresponds to a multiplicative process has two consequences. First, a sample obtained at a range of ages from a multiplicative process (with or without decay) leads to the double Pareto lognormal distribution seen in Fig. 5 (Reed and Jorgensen 2004). In our case, the sample has a uniform range of ages from 1 to R, though with the decay older resolves accumulate votes more slowly than younger ones. A second consequence arises from the decay as resolves become less visible over time. Thus our model provides one mechanism, using information available to users, giving rise to dynamics governed by multiplicative random variation with decay. A similar process arises if the decay is due to any combination of decreasing interest in the content and loss of visibility with age, e.g., as seen in sites such as Digg (Wu and Huberman 2007) with current events stories that become less relevant over time.

Online Estimation

Our model allows estimating parameters for new users, as they vote, and new resolves, as they accumulate votes. In particular, the early history of resolves allows estimating the number of votes a resolve will eventually receive as well as which resolve will likely receive the next vote. Similarly, early user reactions to posted content are accurately predict later popularity of Youtube videos and Digg submissions (Szabo and Huberman 2008). In essence, since content interestingness is largely unchanged in time we can infer its value soon after content is submitted. We can estimate the prediction accuracy by training the prediction algorithm with historic data. The predictions do not consider semantic features of the submitted content (such as title, description, or tags), only initial samples of user activity relating to the content.

Fig. 8 shows estimates of user activity levels from the model described above, as a function of time since the user first voted. Users not only differ considerably in their average activity rates but also in how their interest in the site varies in time. In spite of this temporal variation, early estimates of ρ for active users correlate significantly with their final values. This correlation is 0.41 and 0.62 for estimates after 1 and 7 days of activity, respectively. So while we cannot give definite predictions of future user activity rates from their early experience on the site, we can fairly well distinguish between the more and less active users from their early behavior. On average, user activity rates decline in time, e.g., with the final and early estimates related by $\rho \approx 0.7\rho(7)^{0.8}$ where $\rho(7)$ is the estimate after 7 days.

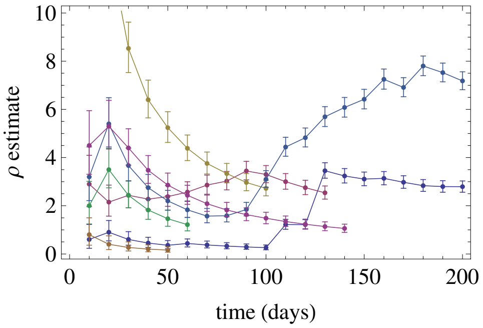

Figure 8: Estimates of ρ values for several users as a function of the time since their first vote. Error bars show the 95% confidence intervals.

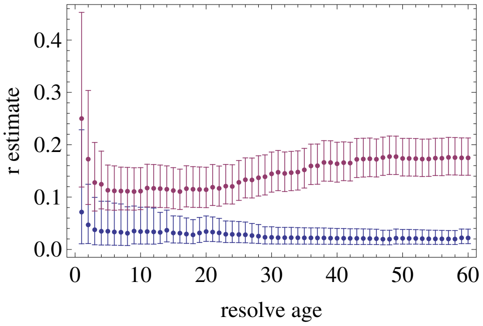

Figure 9: Estimates of r values for two resolves as a function of their age. Error bars show the 95% confidence intervals.

For resolves, Fig. 9 shows how estimates for resolves' r values and their confidence intervals change as more votes are observed. Other resolves show similar behavior. Thus the interestingness of resolves appears to converge in time.

In practice, however, the optimization procedure is computationally costly due to the large number of parameters that grows linearly with the number of resolves in the system. A further requirement of an online algorithm is that it is able to update the model parameters in real time as new users, votes and resolves enter the system. Thus it is not feasible to consider a growing number of resolves with constant resources. Instead we must limit the the number of parameters and thus resolves to be optimized to a constant value. One such approach is to optimize parameters based on the last K active resolves only, and keep the interestingness and aging parameters constant for resolves of age greater than K. This method can also track changes to interestingness and aging in time.

Another incremental approach uses the observation that old resolves, with a long track record of votes, have their interestingness well-estimated. Similarly, the aging function $f(a)$ for small ages is well-estimated from prior experience with many resolves receiving votes at those ages. Conversely, recently introduced resolves have had little time to accumulate votes and $f(a)$ for large ages is poorly estimated due to having little experience in the system with resolves that old. Furthermore, we expect $f(a)$ to change slowly with time, primarily due to how the user interface makes resolves visible to users. For new resolves, the maximum likelihood estimate for the r_j values involves values of $f(a)$ for small ages which are already well-determined from the prior history of the system. So instead of an expensive reevaluation of all the r and f values, we can simply incrementally estimate the r values of new resolves assuming $f(a)$ values for small ages do not change much. Conversely, as new resolves are introduced, the oldest resolves in the system advance to ever larger ages, allowing estimates of $f(a)$ for those ages by assuming the r values of those old resolves do not change much with the introduction of new resolves.

Such estimates of model parameters can be useful guides for improving social web sites, by identifying new users likely to become highly active or content likely to become popular. Since it is possible to estimate the statistical errors given the sample size, one can also perform risk assessment when giving the estimates. Newly posted content with high interestingness, for instance, can be quickly identified and given prominent attention on the online interface.

Discussion

We described several extended distributions resulting from user behavior on Essembly, a web site where users create and rate content as well as form networks. These distributions are common in participatory web sites. From the extended distributions of user behavior we find extremely heterogenous population of users and resolves. We introduce a mechanism describing user behavior based on information available to users, involving a combination of aging and a large variation among people and resolves. We focused on two areas: the wide range in user activity levels and the factors influencing the popularity of user-created content.

We found, first, most users try the online services only briefly, so most activity arises from a relatively small fraction of users who account for the diverse behavior observed. User activity *rates* appear to arise from an underlying multiplicative random process, while activity time is described by a Weibull distribution where users are more likely to continue the longer they have been active. This latter observation contrasts with a simpler model of user activity time as a Poisson process where the probability a user abandons the site is independent of how long the user has already participated. In terms of understanding user behavior, the Weibull distribution raises the significant question of whether the total time users are willing to devote to the site arises from intrinsic heterogeneity in the user population or is mainly due to differing user experiences on the site. In the former case, the decreasing stopping probability with time on the site would arise from an increasing concentration of users with high intrinsic motivation in the population as less dedicated users quit. In the latter case, the behavior arises from positive experiences on the site encouraging continued use. By contrast, the broad distribution of activity rates and their low correlation with activity time suggest activity rates, for however long a user chooses to participate in the site, are mainly due to prior differences among users' interests.

Second, we proposed a model and algorithm that describes and predicts through iterative refinements how the popularity of user-generated content evolves in time, considering both the exposure on the site and inherent interestingness. We found that the exposure content receives depends largely on its recency, and decays with age.

The characteristics of our models apply to other web sites where user participation is self-directed and where content creation and social link formation plays a dominant part in the individual online activities. For example, the Digg and Wikipedia user communities show similar behavior in their activity patterns (Wilkinson 2008).

Consequences of our model include suggestions for identifying user activity level and interesting resolves early in their history. This possibility arises from persistence in voting rates over time, even before content accumulates enough votes to be rated as popular, as is also seen in larger user communities (Szabo and Huberman 2008). Such identification could help promote interesting content on the web site more rapidly, particularly in the case of niche interests. Beyond helping users find interesting content, designs informed by models could also help with derivative applications, such as collaborative filtering or developing trust and reputations, by quickly focusing on the most significant users or items. Such applications raise significant questions of the relevant time scales. That is, observed behavior is noisy, so there is a tradeoff between using a long time to accumulate enough statistics to calibrate the model vs. using a short time to allow responsiveness faster than other proxies for user interest such as popularity.

Our models raise additional questions, such as understanding how the resolve aging function relates to the user interface and changing interests among the user population. Another question is how the wide distributions in user activity and resolve interestingness arise. The lognormal fits suggest underlying multiplicative processes are involved. It would also be interesting to extend the model to identify niche resolves, i.e., resolves of high interest to small subgroups of users but not to the population as a whole. Automatically identifying such subgroups could help people find others with similar interests by supplementing comparisons based on ideological profiles.

A caveat on our results, as with other observational studies of web behavior, is the evidence for mechanisms is based on correlations in observations. While mechanisms proposed here are plausible causal explanations since they rely on information and actions available to users rather than aggregated descriptive variables not known by individual users, intervention experiments would give more confidence in distinguishing correlation from causal relationships. Our model provides testable hypotheses for such experiments. For example, if intrinsic interest in resolves is a major factor in users' selection of resolves, then deliberate changes in the number of votes may change visibility but will not affect interestingness. In that case, we would expect subsequent votes to return to the original trend. Thus one area for experimentation is to determine how users value content on various web sites. For example, if items are valued mainly because others value them (e.g., fashion items and a variety of other economic contexts (Ariely 2008)) then observed votes would *cause* rather than just reflect high value.

In such cases, random initial variations in ratings would be amplified, and show very different results if repeated or tried on separate subgroups of the population. If items all have similar values and differences are mainly due to visibility, e.g., recency or popularity, then we would expect votes due to rank order of votes (e.g., whether item is most popular) rather than absolute number of votes. If items have broad intrinsic value, then voting would show persistence over time and similar outcomes for independent subgroups. It would also be useful to identify aspects of the model that could be tested in small groups, thereby allowing detailed and well-controlled laboratory experiments comparing multiple interventions. Larger scale experiments (Bainbridge 2007; Salganik, Dodds, and Watts 2006) would also be useful to determine the generality of these mechanisms.

Our models incorporate the key features of continual arrival of new users, existing users becoming inactive and a wide range of activity levels among the user population and interest in the content. These features can apply in many contexts. For the distribution of how users rate content (e.g., votes on resolves in Essembly), generalizing to other situations will depend on the origin of perceived value to the users. At one extreme, which seems to apply to Essembly, the items themselves have a wide range of appeal to the user population, leading some content to consistently attract user attention at much higher rates than other content of about the same age. At the other extreme, perceived value could be largely driven by popularity among the users, or subgroups of users, as seen in some cultural products (Salganik, Dodds, and Watts 2006). In rapidly changing situations, e.g., current news events or blog posts, recency is important not only in providing visibility through the system's user interface, but also determining the level of interest. In other situations, the level of interest in the items changes slowly, if at all, as appears to be the case for resolves in Essembly concerning broad political questions such as the benefits of free trade. All these situations can lead to long-tail distributions through a combination of a "rich get richer" multiplicative process and decay with age. But these situations have different underlying causal mechanisms and hence different implications for how changes in the site will affect user behavior. Thus, applying models to the design and evaluation of participatory web sites can benefit from the availability of models relating user behavior to information readily available on the site.

Acknowledgments

We thank Chris Chan and Jimmy Kittiyachavalit of Essembly for their help in accessing the Essembly data. We have benefited from discussions with Michael Brzozowski and Dennis Wilkinson.

References

Anderson, C. 2006. *The Long Tail: Why the Future of Business is Selling Less of More*. Hyperion.

Ariely, D. 2008. *Predictably Irrational: The Hidden Forces That Shape Our Decisions*. NY: Harper Collins.

Bainbridge, W. S. 2007. The scientific research potential of virtual worlds. *Science* 317:472–476.

Boccaletti, S.; Latora, V.; Moreno, Y.; Chavez, M.; and Hwang, D.-U. 2006. Complex networks: Structure and dynamics. *Physics Reports* 424:175–308.

Frisch, U., and Sornette, D. 1997. Extreme deviations and applications. *J. Physics I France* 7:1155–1171.

Glen, A. G.; Leemis, L. M.; and Drew, J. H. 2004. Computing the distribution of the product of two continuous random variables. *Computational Statistics and Data Analysis* 44:451–464.

Guha, R.; Kumar, R.; Raghavan, P.; and Tomkins, A. 2004. Propagation of trust and distrust. In *Proc. of the 13th Intl. World Wide Web Conf. (WWW2004)*, 403–412. New York: ACM.

Hogg, T.; Wilkinson, D. M.; Szabo, G.; and Brzozowski, M. 2008. Multiple relationship types in online communities and social networks. In Lerman, K., et al., eds., *Proc. of the AAAI Symposium on Social Information Processing*, 30–35.

Huberman, B. A., and Adamic, L. A. 1999. Growth dynamics of the World Wide Web. *Nature* 401:131.

Huberman, B. A.; Pirolli, P. L. T.; Pitkow, J. E.; and Lukose, R. M. 1998. Strong regularities in World Wide Web surfing. *Science* 280:95–97.

James, A., and Plank, M. J. 2007. On fitting power laws to ecological data. arxiv.org preprint 0712.0613.

Lam, C. 2004. SNACK: incorporating social network information in automated collaborative filtering. In *Proc. of the 5th ACM Conference on Electronic Commerce (EC'04)*, 254–255. ACM Press.

Lerman, K. 2007a. Social information processing in social news aggregation. arxiv.org preprint cs.cy/0703087.

Lerman, K. 2007b. User participation in social media: Digg study. In *IEEE/WIC/ACM Intl. Conf. on Web Intelligence and Intelligent Agent Technology*, 255–258.

Mitzenmacher, M. 2004. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics* 1:226–251.

Newman, M. E. J. 2005. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics* 46:323–351.

Reed, W. J., and Jorgensen, M. 2004. The double Pareto-lognormal distribution: A new parametric model for size distributions. *Communications in Statistics: Theory and Methods* 33:1733–1753.

Salganik, M. J.; Dodds, P. S.; and Watts, D. J. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *Science* 311:854–856.

Shockley, W. 1957. On the statistics of individual variations of productivity in research laboratories. *Proc. of the IRE* 45:279–290.

Szabo, G., and Huberman, B. A. 2008. Predicting the popularity of online content. Technical report, HP Labs. Available at hpl.hp.com/research/scl/papers/predictions.

Vázquez, A.; Oliveira, J. G.; Dezso, Z.; Goh, K.-I.; Kondor, I.; and Barabasi, A.-L. 2006. Modeling bursts and heavy tails in human dynamics. *Physical Review E* 73:036127.

Vázquez, A. 2003. Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *Physical Review E* 67:056104.

Wilkinson, D. M. 2008. Strong regularities in online peer production. In *Proc. of the 2008 ACM Conference on E-Commerce*, 302–309.

Wu, F., and Huberman, B. A. 2007. Novelty and collective attention. *Proc. of the Natl. Acad. Sci.* 104:17599–17601.