Strong regularities in online peer production

Dennis M. Wilkinson
Social Computing Lab, HP Labs
1501 Page Mill Rd.
Palo Alto, CA
dennis.wilkinson@hp.com

ABSTRACT

Online peer production systems have enabled people to coactively create, share, classify, and rate content on an unprecedented scale. This paper describes strong macroscopic regularities in how people contribute to peer production systems, and shows how these regularities arise from simple dynamical rules. First, it is demonstrated that the probability a person stops contributing varies inversely with the number of contributions he has made. This rule leads to a power law distribution for the number of contributions per person in which a small number of very active users make most of the contributions. The rule also implies that the power law exponent is proportional to the effort required to contribute, as justified by the data. Second, the level of activity per topic is shown to follow a lognormal distribution generated by a stochastic reinforcement mechanism. A small number of very popular topics thus accumulate the vast majority of contributions. These trends are demonstrated to hold across hundreds of millions of contributions to four disparate peer production systems of differing scope, interface style, and purpose.

1. INTRODUCTION

The past decade has seen the emergence of a wide variety of online peer production efforts in which content is created, shared, promoted, and classified by the interrelated actions of a large number of users. Examples include open source software development, collections of wikis (web pages users can edit with a browser), social bookmarking services, news aggregators, and many others. These "coactive" systems now comprise a significant portion of the most visited websites [9] and it is reasonable to assume that they will continue to grow in relevance as Internet use becomes more and more widespread.

Large coactive systems are complex at a microscopic level because there is a high degree of variability in people's decisions to participate and in their reactions to others' contributions. The number of possible interactions is also very large, increasing as the square of the number of participants, and the barrier to interaction online is often lower than in traditional social systems. Nevertheless, as we show, macroscopic regularities can be distinguished given a large enough population and explained in terms of simple individual-level mechanisms. Electronic activity records, being extensive, exhaustive, and easy to analyze, are invaluable for this approach.

Beyond providing interesting descriptions of people's behavior, macroscopic trends in coactive systems are of practical relevance. For example, the basic principle of Internet search is that high quality pages can be differentiated by having accumulated far more visibility and reputation, in the form of incoming links [2]. Another example is the popular success of Wikipedia, which is at least partially due to the correlation between greater user participation and higher article quality [19]. It is rather remarkable that coactivity on such a large scale is able to produce successful results; in many offline applications, result quality plateaus or decreases as the number of collaborators increases past a certain level (e.g. [3, 5]).

Two key challenges in the study of large social systems are to distinguish between general and system-dependent trends, and to provide an explanation for how the trends come about. Empirical regularities which go beyond one particular system or which arise from simple dynamical rules reflect deeply on people's behavior and may be reasonably extended to similar or future instances. A good example of this is the study of social networks, where comparisons of structural properties across a number of disparate networks (e.g. [10]), along with theoretical mechanisms for network formation (e.g. [16]) have combined to provide valuable insight. Other examples include the law of Web surfing [8] and the growth dynamics of the World Wide Web [7].

This paper demonstrates strong macroscopic regularities in four online peer production systems. The regularities are observed in hundreds of millions of contributions made over many years to four systems: Wikipedia, an online encyclopedia anyone with a web browser can edit; Digg, a news aggregator where users vote to identify interesting news stories; Bugzilla, a system for reporting and collaborating to fix errors in large software projects; and Essembly, a forum where users create and vote on politically oriented resolves. While all large, these systems range broadly in scope, size, and purpose, as further discussed below.

The paper presents and examines two fundamental observations: first, that the distribution of levels of user participation is power-law and second, that the distribution of activity per topic is log-normal. We show that these distributions arise from simple rules of participation which illuminate key dynamical properties of peer production. The regularities we observe in these distributions are consistent across the four disparate, independent systems, suggesting their general relevance to the study of coactive participation and collaboration in online peer production.

It is not the goal of this paper to evaluate or guess at the psychological and sociological principles underlying the mechanisms that cause the observed behavior. Rather, it intends to demonstrate the feasibility of a general study of peer production systems and to begin to elicit some of the basic dynamical rules guiding their evolution.

The organization of the paper is as follows. Section 2 describes the peer production systems and our data sets. User participation levels are the subject of section 3, while section 4 discusses the distribution of activity per topic. Section 5 is the summary and conclusion.

2. SYSTEMS AND DATA

The results in this paper were observed in data from four online peer production systems. The data sets from these systems are in all cases exhaustive, in the sense that they extend back to the system's inception and include virtually all contributions by all users. A summary is provided in table 1.

The great variance in focus and scope of the systems analyzed in this paper is a key factor in the generality of the results. Differences in scope are demonstrated in the table. As far as focus, Wikipedia is very broad, Bugzilla is narrow and esoteric, Digg is rather broad but centers on technological news, and Essembly is primarily political in nature. It is reasonable to assume that the population of contributors to each system represents a different cross section of Internet users.

System	time span of data	users	topics	contribs.
Wikipedia	6 yrs, 10 mos	5.07 M	1.50 M	50.0 M
Bugzilla	9 yrs, 7 mos	111 k	357 k	3.08 M
Digg	3 yrs, 0 mos	1.05 M	3.57 M	105 M
Essembly	1 yr, 4 mos	12.4 k	24.9 k	1.31 M

Table 1: Data sets in this paper. "Topics" refers to articles in Wikipedia, bugs in Bugzilla, stories in Digg, and resolves in Essembly. "Contributions" refers to non-robot edits in Wikipedia, comments in Bugzilla, "diggs" or votes in Digg, and votes in Essembly.

Wikipedia [1] is the online encyclopedia which any user can edit. It consists of a large number of articles (as of this writing, over 9 million [4]) in wiki format, that is, web pages users can edit using a web browser. Wikipedia users can and often do submit multiple edits to a single page. All previous article versions are cached and users can review these as well as exchange comments on the article's dedicated talkpage. When editing, people are encouraged to follow a code of principles and guidelines, and in the worst cases of misuse, volunteer administrators may step in and ban a particular editor for a short time. Users can locate Wikipedia articles using a search function, and the articles are also hyperlinked together when related terms appear in the text.

Our data set contains user ID, article ID and timestamp for all the edits made to the English language Wikipedia between its inception in January 2001 and November 2, 2006. We processed the data to exclude disambiguation and redirect articles, as well as the 5.2 million edits made by robots, as described in [19].

[1] www.wikipedia.org

Essembly [2] is an open online community where members propose and vote on politically oriented resolves, post comments, and form friendships, alliances and anti-alliances ("nemesis links"). The site's welcome page states that its goal is to allow users to "connect with one another, engage in constructive discussion, and organize to take action," although personal experience suggests that voting and commenting on resolves is the dominant activity. Any user can write and upload a resolve using a web browser. Many of the resolves are political in nature, while others are casual. Voting is done on a four point scale ranging from strongly agree to strongly disagree, and one's votes are visible to neighbors in the social networks. Only one vote is allowed per resolve. Within Essembly, multiple mechanisms exist for users to learn about new resolves, including lists of recent popular or controversial resolves and votes within users' social and preference networks, none of which is particularly dominant [6].

Our data set contains randomized user ID, randomized resolve ID and timestamp for all resolve submissions and votes cast between Essembly's inception in August 2005 and December 12, 2006.

Bugzilla is an online service for reporting errors and collaborating to fix them in software development efforts. Any large software project can have its own bugzilla; our data comes from the Mozilla Bugzilla[3]. (Mozilla is an open-source suite of Internet tools including a web browser, email client, and many others, and is a large project involving many thousands of developers.) Within Bugzilla, each reported bug has its own page where users can post detailed information, examples, patches and fixes, and exchange comments. The comments typically discuss technical matters and users may comment multiple times on a single bug. A comment almost always accompanies a patch, fix or other form of resolution. Bugzilla is equipped with a search function to help users find bugs, and lists of related or dependent bugs exist for some bugs.

Our data set contains randomized user ID and bug ID for the 3.08 million comments posted under the first 357,351 reported Mozilla bugs, from April 1998 through November 22, 2006.[4]

Digg [5] is a social news aggregator where users submit and vote for, or "digg," online news stories they find interesting. A Digg vote can only be positive, and indicates that the users finds the story interesting. Only one vote is allowed per story. Any user may submit a story, in the form of a URL link, provided it has not been previously submitted. Fifteen popular recent stories appear on the front page, according to a proprietary algorithm, and beyond this users must use a search function to find stories. The popular stories are updated on the time scale of minutes.

Our data set consists of user IDs, story IDs and timestamps for all the story submissions and votes cast between Digg's inception in December 2004 and December 5, 2007 [6].

[2] www.essembly.com

[3] https://bugzilla.mozilla.org/

[4] This figure excludes some 3500 bugs which we required special authorization to access, most likely because of security concerns, and include 54 older bugs imported from Netscape bug lists.

[5] www.digg.com

[6] It appears that approximately 1% of the digg contributions, roughly randomly distributed in time and by story, were removed

3. USER PARTICIPATION

In every social unit, there is a range in the amount of participation by different members, from a dedicated core group to a periphery of occasional or one-time participants. The distribution of user participation in social systems is of practical relevance to the understanding of these communities and how they evolve. As we show, in the systems we are considering, participation follows a power law distribution in which a small number of very active users account for most of the activity. This form is general to the extent that these systems are representative of online peer production. A heavy tail trend was previously noted in chat room posts [18], but the distribution was not formally studied or extended to other online communities.

To better compare across different systems, we will initially consider only those users who are inactive, meaning that they have not contributed for several months or more. Contribution counts for inactive users are "final," in the sense that these users have almost certainly made a decision, conscious or incidental, to stop participating in the peer production system. The decision to stop is a key focus of this section. It is less meaningful to compare contribution counts for active users across various systems, since it depends on when the observation is made in terms of the system's "life cycle" of growth or decline. Nevertheless, because of its practical relevance, we do present observations for active users and a short discussion in subsection 3.4.

The central results of this section are as follows:

1. Inactive users' "final" contribution counts follow a power law;
2. The power law arises because the probability a user quits after making k contributions is equal to $(\alpha - 1)/k$, where α is the power law exponent for the system;
3. The power law exponent is strongly related to the system's barrier to contribution, in light of the $(\alpha - 1)/k$ rule and as justified by the data.
4. The distribution of contributions for all users, active and inactive, is also power law with a smaller exponent than for the power law for inactive users.

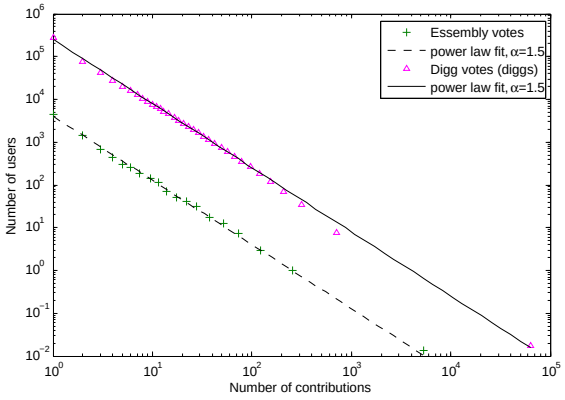
3.1 Observations

For the systems under consideration in this paper, we measured participation in the following ways. For Wikipedia, we counted the number of non-robot edits[7] made by each user. In Bugzilla, we counted the number of non-robot comments, including those accompanying patches or other resolutions, posted by each user. In Digg and Essembly, we measured participation in two ways: first, by counting the number of votes (in Digg known as "diggs") per user, and second, by counting the number of stories or resolves submitted.

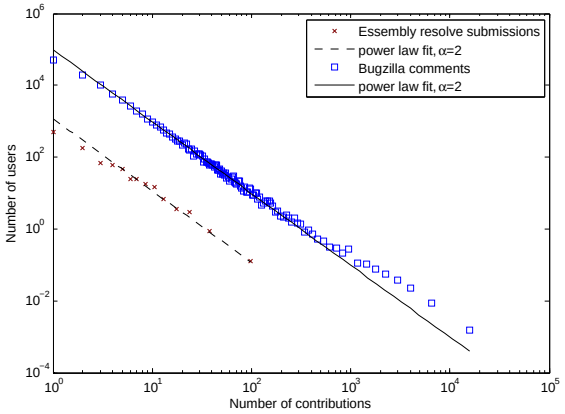
Inactive users were defined as those who had not contributed for 3 months (Digg and Essembly) or 6 months (Wikipedia and Bugzilla)

from the data set before we obtained it. This is possibly because a few users asked the website administrators to delete their actions from the records [15].

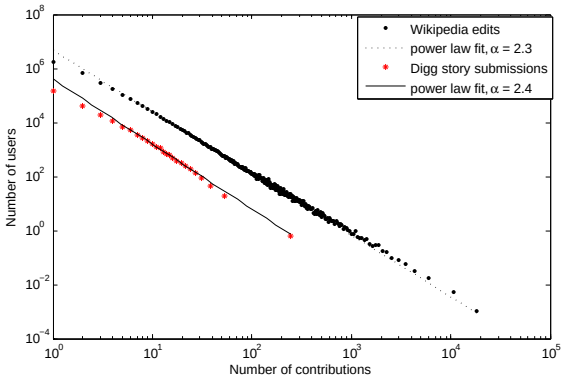
[7]Robot edits were identified as being made by Wikipedia-registered robots, and also as all edits made within 10 seconds of the previous edit. Any actual human edits excluded by this cutoff were not likely to have been significant contributions of content.



(a) Digg and Essembly votes



(b) Bugzilla comments and Essembly resolve submissions



(c) Wikipedia edits and Digg story submissions

Figure 1: Empirical probability density functions for the "final" number of contributions per user for inactive users. The best power law fit is included for comparison.

prior to the latest day for which we had data. It was necessary to use 3 months for Digg and Essembly because the time span of our data was shorter for these systems. Inactive users made up 71% of Wikipedia editors, and 95% of Bugzilla commentors, 61% of Digg voters, 56% of Digg story submitters, 83% of Essembly voters and 53% of Essembly story submitters.

Figure 1 demonstrates the distribution of final contribution counts for all six modes of contribution considered in this paper. The data are plotted on a loglog scale and a power law fit is included for comparison. In contrast to most studies of empirical power law distributions, our focus is primarily not the tail (where data are scarce) but the central part of the distributions. An equal count binning procedure was used where the bin size was proportional to the total number of users in the system. This procedure produces a number density curve, which is equivalent to a probability density function multiplied by the total number of users.

The descriptive accuracy of the power law is clear from the figure. In addition, statistical tests suggest that the power law is generative, except for the lowest values of k, for all contribution types except for Digg submissions. Table 2 shows the p-values obtained using likelihood ratio G-tests[8]. The slight deviation at the high end of some distributions is not of particular interest for this paper because of the small number of counts in this range.

Contribution type	α	p-value	min. k
Essembly votes	1.47	0.59	3
Digg votes	1.53	0.64	15
Bugzilla comments	1.98	0.74	5
Essembly submissions	2.02	0.25	7
Wikipedia edits	2.28	0.69	10
Digg submissions	2.4	0.04	15

Table 2: p-values for power law fit to data. α is the power law exponent which achieved the given p value, and "min. k" means that the power law only fit the data for users making k or more contributions.

3.2 Participation "momentum"

The power law's excellent description of the true distributions over their entire range suggests the following interpretation in terms of when people stop participating. Mathematically, the power law means that the number of people $N(k)$ who have made k contributions is given by

$$N(k) = Ck^{-\alpha},$$

where C is a constant determined by the total number of users in the system[9].

The probability that a user stops after his kth contribution is equal to the number of users contributing exactly k times divided by the number of users contributing k or more times:

$$P(\text{stop after } k) = \frac{Ck^{-\alpha}}{C\sum_{b=0}^{\infty}(k+b)^{-\alpha}} = \frac{1}{\sum_{b=0}^{\infty}(1+b/k)^{-\alpha}}. \quad (1)$$

In the large k limit,

$$\begin{aligned}\frac{1}{k}\sum_{b=0}^{\infty}\left(1+\frac{b}{k}\right)^{-\alpha} &= \int_0^{\infty}(1+x)^{-\alpha}dx + O(1/k) \\ &= \frac{1}{\alpha-1} + O(1/k)\end{aligned}$$

where we have used Riemann integration with step size $1/k$. In fact, since the maximum slope of the function $(1+x)^{-\alpha}$ on $(0,\infty)$ is $-\alpha$, the error term is bounded above by $\alpha/2k$ [1]. Returning to equation 1, we have that

$$P(\text{stop after } k) = \frac{\alpha-1}{k} + O(1/k^2) \quad (2)$$

where the error term is bounded above by $\alpha(\alpha-1)^2/2k^2$ and is thus very small for k as small as 5 or 10 for the values of α we observe.

Equation 2 indicates that people have a "momentum" associated with their participation, such that their likelihood of quitting after k of contributions decreases inversely with k. This rule holds for any power law, independent of the value of the exponent. The rule is confirmed in figure 2, where the proportion of users quitting after k contributions is shown for Wikipedia edits and Essembly votes. Compare the data to the fitted lines of $(\alpha-1)/k$ where the values of $\alpha = 1.5$ for Essembly and $\alpha = 2.3$ for Wikipedia were previously observed in figure 1. A similar fit is observed for the other forms of contribution under discussion.

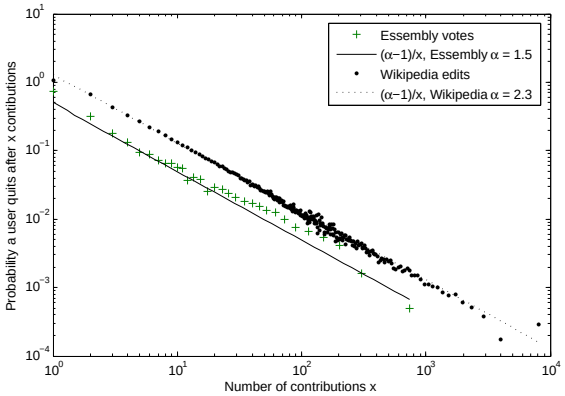


Figure 2: Momentum law that a user quits after x edits with probability $(\alpha-1)/x$. We note that the fitted curves have no free parameters; the values $\alpha = 1.5$ for Essembly and $\alpha = 2.3$ for Wikipedia were taken from the observations of figure 1.

3.3 Interpretation of the power law exponent in terms of effort required to contribute

The previous discussion suggests a straightforward interpretation of the meaning of the power law exponent α. In equation 2 for the probability a user quits contributing, larger values of α indicate that at every opportunity, a contributor is more likely to quit. When the effort required to to contribute is higher, we thus expect a larger value of α.

[8]This test is appropriate because we are more concerned with the central part of the distribution than the tail, so that bin size does not strongly affect the result.

[9]$N(k)$ is in fact a number density function, a distinction which only matters when k so large that $N(k)$ is fractional; in this case, $N(k)$ should be regarded as the expected or average number of users to have made k contributors over a large number of (hypothetical) systems.

Voting in Digg and Essembly can be done quickly with little personal investment [10]. More effort is required in the submission of a new Digg story or making a Wikipedia edit, where the user is required to do some background search and then formulate his submission. We therefore expect to find a higher value of α for Wikipedia edits or Digg submissions than for Digg and Essembly votes.

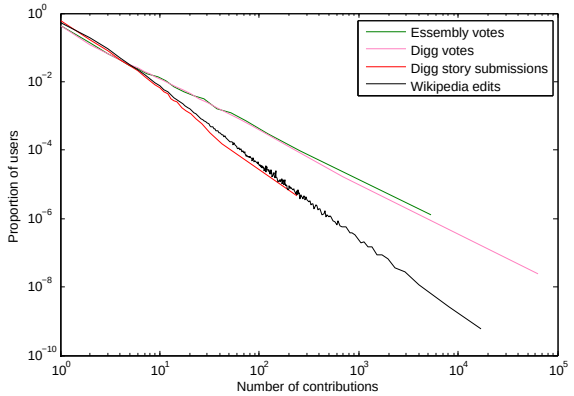


Figure 3: Empirical probability density functions for the "final" number of contributions per user for inactive users.

This expectation is confirmed by the data, as shown in figure 3 and table 2. In this figure, we have produced an empirical probability density function for each system by dividing each user's counts by the total for that system and binning as before.

It is striking that the power law exponents are so similar for Digg and Essembly voting, and for Wikipedia edits and Digg submissions. This suggests that the barrier to participation is the dominant element in determining α and thus the rate of participation dropoff, which has obvious implications for system design. We also note that Bugzilla comments and Essembly submission with $\alpha \approx 2$ provide an intermediate case. These contributions involve a highly variable amount of effort, ranging from very little effort for a casual Essembly resolve or response to a colleague's Bugzilla comment, to a great deal of effort for bug fixes or Essembly political statements.

3.4 Contribution counts for all users

From a practical standpoint, it may be of interest to consider the distribution of counts by all users, active and inactive, at a given time. The distributions of counts for all users is shown in figure 4, where a power law is still quite descriptive. The figure demonstrates that when active users are taken into account, we still observe a power law form for the distributions.

The best fit exponents for these distributions, as well as for Essembly submissions, are shown in table 3. We still observe a strong correlation between the power law exponent and the system's barrier to contribution. As we might expect, the exponents are smaller than the corresponding exponents when only inactive users are considered, because the proportion of active users in the system increases with the number of contributions. The relation between the exponents for inactive users and for all users depends on a number of

[10]This quick approach is evidenced by (for example) the rapid accumulation of votes in both systems immediately following the appearance of a resolve or story [15].

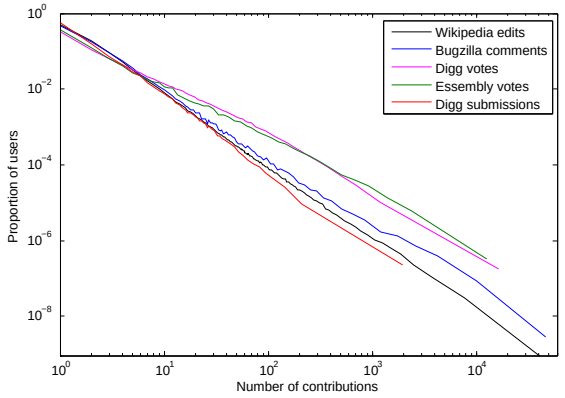


Figure 4: Empirical probability distribution functions for the number of contributions per user, for both inactive and active users, as of the latest date for which we had data.

factors, including the rate at which contributions are made and the rate at which new users appear in the system, and is beyond the scope of this paper.

Contribution type	α
Essembly votes	1.38
Digg votes	1.35
Bugzilla comments	1.92
Essembly submissions	1.78
Wikipedia edits	1.96
Digg submissions	2.00

Table 3: Best fit power law exponent α for the distributions of contributions by all users, active and inactive

As more and more users in the system become inactive, the distribution of user contributions tends toward that of the "final" distribution counts we observed for inactive users only, meaning that the power law exponent increases. This is evident in the data of tables 2 and 3. The power law exponent for inactive users is almost identical to that for all users in Bugzilla, where 95% of the users are inactive. In contrast, the exponent for inactive users is a decidedly greater than the exponent for all users for Wikipedia editing, Digg voting and submissions, and Essembly submissions, where respectively 71%, 61%, 56% and 53% of users are inactive. Although it is somewhat of a tautology, we point out here that the tendency of the power law exponent to increase as a higher proportion of users become inactive means that an ever larger percentage of the contributions are made by very active users (as the influx of new users slows).

4. TOPIC ACTIVITY

We now turn from the question of number of contributions per user to the number of contributions per story or topic. This subject is of significant practical importance, as demonstrated by the examples of Google search and Wikipedia quality we mentioned in the introduction. Just as for user participation levels, the distribution has a heavy tail of very popular topics which attract a disproportionately large percentage of participation and interest. In this case, however, the exact form is lognormal, not power law, implying a different generative mechanism.

In this section, contributions are counted as before, and "topics" refers to Wikipedia articles, Essembly resolves, and Digg stories. To measure of the level of activity on a topic, the procedure was to simply count the number of contributions to it. For Wikipedia, this metric was shown to correlate strongly to page views [14]. The Bugzilla data are not included here because the reinforcement mechanism of this section is not applicable to the Bugzilla process.

The central results of this section are as follows:

1. The distribution of contributions per topic, among topics of the same age, is lognormal;
2. Where novelty decay is not a factor, the lognormal mean and variance depend linearly on time;
3. These observations are explained by a multiplicative mechanism in which contribution reinforces visibility and popularity.

We first describe the theory behind multiplicative reinforcement and then present our observations.

4.1 Multiplicative reinforcement

Consider the number of new edits to a Wikipedia article, or votes to an Essembly resolve or Digg story, made between time t and time $t + dt$, an interval of minutes or hours. Because of the complicated nature of the system, this number will vary a lot depending on the time period and topic. However, the overall average amount of new activity will be directly related to the visibility or popularity of the topic.

We account for the effect of coaction in the system in the simplest possible way, by assuming that contributions to a topic increases its popularity or visibility by some constant amount, on average, with deviations away from the average absorbed into a noise term. The number of contributions to a given topic will thus be proportional to the number of previous contributions, and the dynamics of the system can be expressed simply as:

$$dN_t = [\mu + \sigma dB_t]N_t dt. \qquad (3)$$

In this equation, N_t is the number of contributions on the topic up until time t; dN_t is the amount of new activity between t and $t + dt$ for some suitably small dt; μ is the average rate of contribution, independent of topic or time; and σB_t is a stochastic Wiener process whose variance is $\sigma^2 t$. That is, dB_t are i.i.d noise terms which embody the vagaries of human behavior, the varying effect that one person's contribution has on other people's participation, and the varying effect each contribution has on topic popularity.

For Wikipedia and Essembly, this equation is sufficient to describe the dynamics. For Digg, it must be modified by introducing a discount factor to account for the decay in novelty of news stories over time [20]. In Digg, the basic equation is thus

$$dN_t = r(t)[a + \xi_t]N_t dt.$$

where $r(t)$ is a monotonically decreasing function of age. Even with the novelty factor, the final distribution of votes per story can be shown to follow a lognormal distribution, but the age dependence is more complex. It is also important to mention that this mechanism only functions in Digg for stories which are shown on the front page, because the site interface so heavily favors these in terms of visibility.

Equation 3 is a stochastic differential equation whose solution $N(t)$ is a probability density function, meaning that the exact number of contributions to a given topic at a given time can take on a range of values. In light of the bursty nature of contributions, we adopt a Stratonovich interpretation[11] and the solution to eq. 3 is the probability density function

$$P[N(t)] = \frac{1}{N\sqrt{2\pi}\sqrt{s^2 t}} \exp\left[-\frac{(\log N - \mu t)^2}{2(\sigma^2 t)}\right], \qquad (4)$$

where again σ^2 is the variance of the stochastic process and μ is the average rate of accumulation of edits or votes [11]. This equation describes a lognormal distribution whose parameters depend linearly on the age t of the topic. Note that μt and $\sigma^2 t$ represent the mean and variance, respectively, of the log of the data, and are thus related to but not equal to the distribution mean and variance.

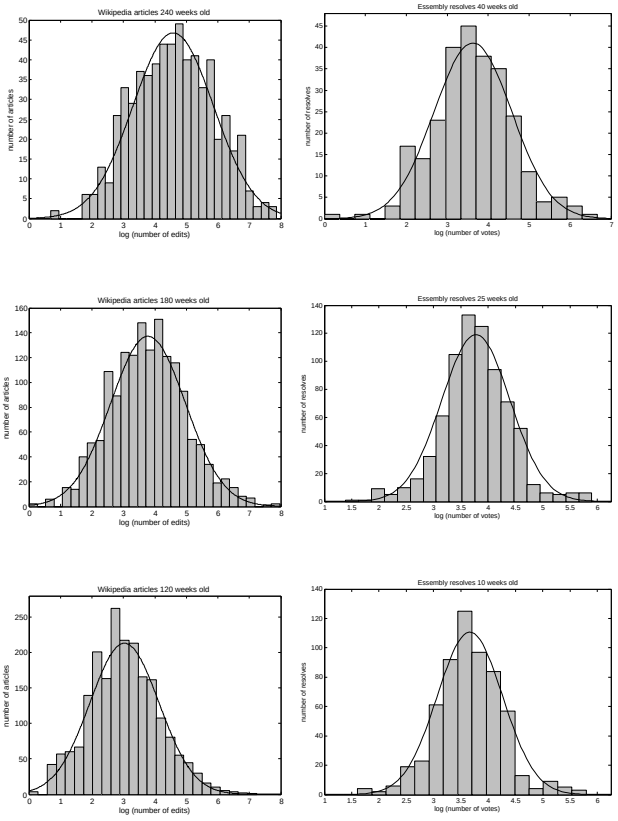
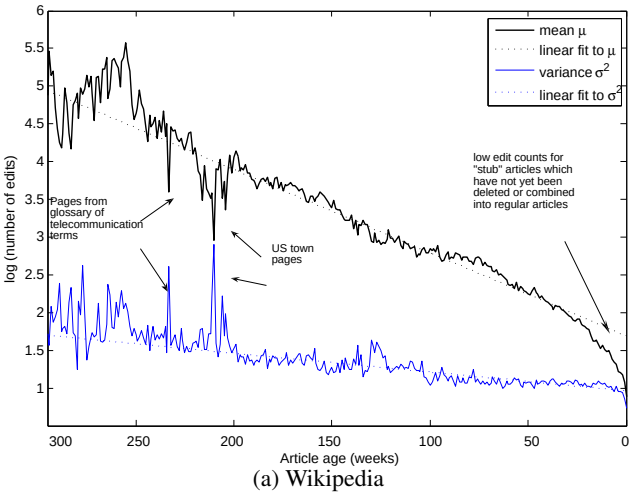


Figure 5: Distributions of the logarithm of the number of Wikipedia edits and Essembly votes for several articles or resolves within several time slices: Wikipedia articles ages 240 weeks (top left), 180 weeks (middle left), and 120 weeks (bottom left); and Essembly resolves ages 40 weeks (top right), 25 weeks (middle right), and 10 weeks (bottom right). Since the number of participations is lognormally distributed, the logarithm is normally distributed. The best fit normal curve is included for comparison.

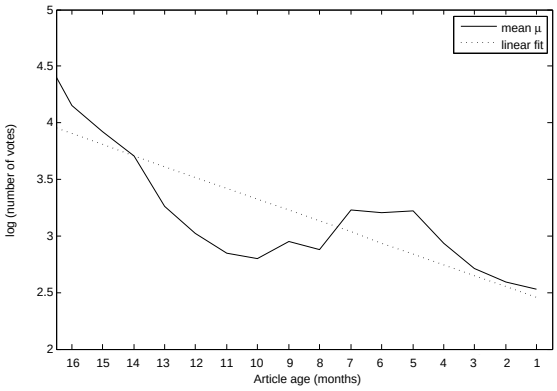
[11] In practice, the Ito interpretation would yield almost the same result because $\sigma^2 \ll \mu$ for the systems we have studied.

4.2 Observations

The model described above predicts that the distribution of contributions per topic will be lognormal for topics of the same age, and that when novelty decay is not a factor, the lognormal parameters μ and σ^2 will vary linearly in time. These predictions are confirmed by the data. A log-likelihood ratio test on the Wikipedia data shows that 47.8 % of the time slices have a p-value greater than 0.5, for a lognormal distribution with the empirical μ and σ^2. A similar test for Essembly, shows that 45.6 % of the time slices have a p-value greater than 0.5. In Digg, statistical tests likewise confirmed the lognormal distribution [20]. The lognormal form of the distribution of contributions per topic is demonstrated in figure 5 for several time slices from Wikipedia and Essembly.



(a) Wikipedia



(b) Essembly

Figure 6: Evolution of the parameter μ, the mean of the logarithm of edit or vote counts, for Wikipedia articles and Essembly resolves. For Wikipedia edits, the evolution of the variance σ^2 is also included. The linear best fit line is included for comparison.

The time dependence of the distribution parameters μ and σ^2 with article or resolve age provides another confirmation of the accuracy of equation 4. The linear dependence of μ, which is the mean of the logarithm of participation counts, with topic age in Wikipedia and Essembly is demonstrated in figures 6. The dependence of the variance σ^2 is also included for Wikipedia. For Wikipedia, occasional large deviations from the pattern are noted and explained in the figure. For Essembly, the number of data are not large enough to demonstrate the trend as clearly; compare the sample variability with the first 50 or so weeks of Wikipedia data (the numbers of data points in these time slices are similar). In Digg, as previously mentioned, the time dependence was more complex because of the decay of novelty, but observed values of the parameters μ and σ^2 were found to have the correct time dependence [20].

Significantly, the role of the system interface for reinforcing popularity is very different for Digg, Essembly, and Wikipedia. In Digg, reinforcement is explicit because the popular stories are prominently featured on the main page. Essembly users can identify new resolves to vote on via overall popularity as well as a number of other mechanisms including their social network and traditional keyword search. Wikipedia has no explicit mechanism in its interface for contributions to reinforce popularity. That multiplicative reinforcement exists in all three systems highlights the importance of social mechanisms in peer production and has important implications for interface design.

The heavy-tail nature of the lognormal distribution means that a small number of topics or stories will attract the vast majority of contributions. It is worth noting that a different model, where topics begin with a varying degree of "inherent" popularity or general interest and then accumulate contributions without a reinforcement mechanism, fails to explain the lognormal distribution of contributions. The observation that popularity reinforcement in the form of contribution has been key part of the evolution of three disparate peer production systems is of practical and theoretical relevance [13].

5. SUMMARY AND CONCLUSION

The main theme of this paper was that disparate forms of online peer production share common macroscopic properties which can be explained by simple dynamical mechanisms. We presented observations from four disparate online social systems: Wikipedia, Digg, Bugzilla and Essembly. Our results are general to the extent that these systems are representative of online peer production. It is hoped that this paper will represent a first step toward understanding the dynamics of peer production systems in which a large number of people coactively create, rate and share content.

First, it was shown that user participation levels in all four systems are well-described by a power law, in which a few very active users account for most of the contributions. The power law arose because there is a "momentum" associated with participation such that the probability of quitting is inversely proportional to the number of previous contributions. The power law exponent was shown to correspond clearly to the effort required to contribute, with higher exponents in systems where more effort is required. A striking similarity was observed in the exponent α in systems requiring similar effort to contribute: $\alpha \approx 2.35$ for Wikipedia edits and Digg submissions, while $\alpha \approx 1.5$ for Digg and Essembly voting. This suggests that the user participation distribution is primarily dependent only on the participation momentum rule and the system's barrier to contribution.

Next, we showed that the distribution of contributions per topic is lognormal because of a multiplicative reinforcement mechanism in which contributions increase popularity. This explains the propensity of a few very visible popular topics to dominate the total activity in coactive systems. It is rather remarkable that the many

forms of variation at the individual level of these systems can be accounted for with such a simple stochastic model. It is also worth reiterating that the mechanism functions in all three systems even though their interfaces favor popularity reinforcement to greatly different degrees.

The observed regularities are of practical relevance to the understanding of online peer production. Governed by simple mechanisms and consistent across a variety of systems, these regularities provides a useful tool for estimation and comparison of metrics such as the barrier to system participation or the rate at which topics accumulate popularity.

The heavy-tail nature of the distributions of contributions per user and per topic also highlights some of the difficulties in predicting which peer production efforts will attain huge size or widespread popularity. For example, the number of contributions made by the most prolific users is centrally important to the total number of contributions: because $\int k^{1-\alpha} dk$ is divergent for $\alpha < 2$, the high end cutoff must be used. However, there is a great deal of variance in the high end cutoff for both the reinforcement mechanism $dN = [\mu + \sigma dB]Ndt$ and the contribution momentum mechanism $P(\text{quit after } k) = (\alpha - 1)/k$.

To put it more plainly, these systems depend a great deal on the very heavy contributors and very popular topics and it is difficult to make predictions about these things, even if the barrier to contribution or total number of contributors can be approximately known. For example, it is reasonable to assume that the outlook or philosophy of the very dedicated or prolific users will have a strong effect on the system, both by their contribution and their social interactions, which goes beyond any quantitative measure of prediction. It has been argued that this effect, more than any other, is responsible for the success of Wikipedia, for example [12].

As a final note, this paper illustrates the importance of large data sets in the study of coactive phenomena. For example, the nearly 25,000 resolves and 12,500 users of Essembly were barely enough to detect the time-dependence in the distribution of topic popularities. Access to electronic records of online activity is thus essential to progress in this area, and it can only be assured if privacy continues to be respected completely as the scientific community has done to date [17].

Acknowledgments The author thanks Gabor Szabo, Mike Brzozowski, and Travis Kriplean for their help processing the data; Bernardo Huberman, Fang Wu, and Gabor Szabo for helpful conversations; Chris Chan and Jimmy Kittiyachavalit of Essembly for their help in preparing and providing the Essembly data; and the Digg development team for allowing API access to their data.

6. REFERENCES

[1] M. Abramowitz and I. Stegun. *Handbook of Mathematical Functions*. Dover, New York, 1964.

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual search engine. *Computer Networks and ISDN Systems*, 30:107–117, 1998.

[3] F. Brooks. *The Mythical Man-month*. Addison-Wesley, Reading, Mass., 1975.

[4] Wikimedia Foundation. http://meta.wikimedia.org/wiki/List_of_Wikipedias, accessed 11/23/2007.

[5] J. R. Galbraith. *Organizational Design*. Addison-Wesley, Reading, Mass., 1977.

[6] T. Hogg, D. M. Wilkinson, G. Szabo, and M. Brzozowski. Multiple relationship types in online communities and social networks. 2008. to appear in Proc. AAAI Conf. on Social Information Processing, 2008.

[7] B. A. Huberman and L. A. Adamic. Growth dynamics of the World Wide Web. *Nature*, 399:130, 1999.

[8] B. A. Huberman, P. Pirolli, J. E. Pitkow, and R. M. Lukose. Strong regularities in World Wide Web surfing. *Science*, 280(5360):95–97, 1998.

[9] Alexa Internet Inc. http://www.alexa.com/site/ds/top_500, accessed 2/5/2007.

[10] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

[11] B. K. Øksendal. *Stochastic Differential Equations: an Introduction with Applications*. Springer, Berlin, 6th edition, 2003.

[12] D. Riehle. How and why Wikipedia works: an interview. In *Proc. ACM Wikisym*, 2005.

[13] M. J. Salganik, P. S. Dodds, and D. J. Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762):854–865, 2006.

[14] A. Spoerri. What is popular on Wikipedia and why? *First Monday*, 12(4), 2007.

[15] G. Szabo and K. Bimpikis. personal communications.

[16] R. Toivonen, J.-P. Onnela, J. Saramäki, Jörkki Hyvönen, and K. Kaski. A model for social networks. *Physica A*, 371(2):851–860, 2006.

[17] D. J. Watts. A twenty-first century science. *Nature*, 445:489, 2007.

[18] S. Whittaker, L. Terveen, W. Hill, and L. Cherny. The dynamics of mass interaction. In *CSCW '98: Proceedings of the 1998 ACM conference on Computer supported cooperative work*, pages 257–264, New York, NY, USA, 1998. ACM.

[19] D. Wilkinson and B. Huberman. Assessing the value of cooperation in Wikipedia. *First Monday*, 12, 2007.

[20] F. Wu and B. Huberman. Novelty and collective attention. *Proc. Natl. Acad. Sci. USA*, 105:17599, 2007.