A Model for Information Growth in Collective Wisdom Processes

SANMAY DAS, Rensselaer Polytechnic Institute
MALIK MAGDON-ISMAIL, Rensselaer Polytechnic Institute

Collaborative media like wikis have become enormously successful venues for information creation. Articles accrue information through the asynchronous editing of users who arrive both seeking information and possibly able to contribute information. Most articles stabilize to high quality, trusted sources of information representing the collective wisdom of all the users who edited the article. We propose a model for *information growth* which relies on two main observations: (i) as an article's quality improves, it attracts visitors at a faster rate (a rich get richer phenomenon); and, simultaneously, (ii) the chances that a new visitor will improve the article drops (there is only so much that can be said about a particular topic). Our model is able to reproduce many features of the edit dynamics observed on Wikipedia; in particular, it captures the observed rise in the edit rate, followed by $1/t$ decay. Despite differences in the media, we also document similar features in the comment rates for a segment of the LiveJournal blogosphere.

Categories and Subject Descriptors: J.4 [**Social and Behavioral Sciences**]: Economics; J.4 [**Social and Behavioral Sciences**]: Sociology; H.4 [**Information Systems Applications**]: Miscellaneous

General Terms: Algorithms, Human Factors, Measurement, Theory

Additional Key Words and Phrases: Collective intelligence, social networks, dynamical systems

ACM Reference Format:
Das, S. and Magdon-Ismail, M. 2011. A Model for Information Growth in Collective Wisdom Processes. ACM Trans. Knowl. Discov. Data. 0, 0, Article 0 (0), 10 pages.
DOI = 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

1. INTRODUCTION

Wikipedia has become a trusted source of information for most web users. Independent studies have verified that Wikipedia articles are of comparable quality to the Encyclopedia Britannica [Giles 2005]. This paper models the dynamics of information growth in "collective intelligence" settings like Wikipedia. We define a *collective wisdom process* (CWP) as a process in which users asynchronously contribute information on a particular topic. Participants in a CWP can be contributors or consumers (or both) of information.

A preliminary version of this paper appears in the *Proceedings of the ACM Conference on Electronic Commerce* [Das and Magdon-Ismail 2010].
This work is supported by a National Science Foundation CAREER award (IIS-0952918), by NSF grants IIS-0621303, IIS-0522672, IIS-0324947, CNS-0323324, IIS-0634875, by U.S. ONR Contract N00014-06-1-0466, and by US DHS through ONR grant N00014-07-1-0150 to Rutgers University. This research is continuing through participation in the Network Science Collaborative Technology Alliance sponsored by the U.S. Army Research Laboratory under Agreement Number W911NF-09-2-0053. The content of this paper does not necessarily reflect the position or policy of the U.S. Government, no official endorsement should be inferred or implied.
Author's addresses: S. Das and M. Magdon-Ismail, Department of Computer Science, Rensselaer Polytechnic Institute, {sanmay,magdon}@cs.rpi.edu.

DOI 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

Scientific studies of Wikipedia and other social media settings (like the Blogosphere) have focused on growth (addition of new articles) [Capocci et al. 2006; Spinellis and Louridas 2008], on macroscopic properties of communication dynamics [Silva et al. 2009], and on network modeling of the implied social interactions [Kumar et al. 2004]. Recently there has also been some focus on what makes content popular [Wu and Huberman 2007; Spoerri 2007]. Another direction of research has studied the sociological implications of these new media, examining the editing behavior of users [Kittur et al. 2007] and the emergence of bureaucracies in Wikipedia [Butler et al. 2008]. While this paper is related to the entire literature on the growth of networks, both general and social [Barabasi and Albert 1999; Leskovec et al. 2008], it differs in its focus on the dynamics of information rather than on the dynamics of user arrivals and departures.

We note also two other directions of related work: much research on human, social and economics dynamics has focused on finding distributions that are a good fit to data. For example, the work of Wilkinson and Huberman [2007] which we discuss in detail below, focuses on fitting Wikipedia edit data to the log-normal distribution. The work of Vaz de Melo et al. [2010] is a nice example of producing well-motivated new distributional models for fitting data gathered from observations of human behavior (a truncated version of the log-logistic distribution fits call duration data well, and can replicate power-law tails in a manner that promises to be useful in other domains). Our focus in this paper is, however, on providing a generative model that explains stylized facts well, rather than best-fits after the fact. In this sense our work is close in spirit to that of Götz et al. [2009], who are among the first to provide a combined temporal and topological model of posting behavior in the context of the blogosphere.

Wikis are mechanisms for sharing knowledge, beliefs, and opinions. They provide a unique opportunity to understand the dynamics of collective wisdom, and in order to do this it is important to focus on the dynamics of the growth of individual articles. In this paper we focus on highly edited wikis from Wikipedia. Wilkinson and Huberman [2007] have shown that highly edited articles tend to be of higher quality on Wikipedia, and we confirm that the most visited Wikipedia pages are also heavily edited. These heavily edited pages form the core of the content for which Wikipedia is most well-known and used. Although the analysis of traffic that goes to the so-called "long tail" of less edited, significantly less popular pages, is independently interesting, we do not focus on it here.

Wilkinson and Huberman [WH] may have been the first to study dynamics in the Wikipedia context. They propose a "rich get richer" stochastic geometric growth model in which articles accrue edits at a rate proportional to the number of edits already received. In this model, letting $n(t)$ denote the number of edits to an article, the number of new edits over a period Δt is given by $\Delta n(t) = (a + \xi(t))n(t)$, where a is the average edit rate and $\xi(t)$ are independent zero-mean random variables which account for random fluctuations in the edit rate. A snapshot of all pages which have been alive for the same amount of time would yield a lognormal edit distribution under this model, and WH take the existence of such a distribution in the data as evidence for their model.

One consequence of simple rich-get-richer models, like the WH model, and others that study wikis and blogs as analogous to network growth processes (NGPs) such as the growth of the WWW [Barabasi and Albert 1999] or the Internet [Faloutsos et al. 1999], is that the total number of edits on a given wiki article or blog post should continue to increase over time. However, CWPs are fundamentally different from NGPs in that they are primarily information processes. There is only a finite amount of information about a given topic, so we would expect wiki pages and comments on blog posts to eventually stabilize to a state that reflects the collective wisdom on a topic. The WH model can be modified to account for this by fitting time-dependent location and scale

parameters for the log-normal distribution, but does not focus on explicitly modeling the saturation of information.

We propose a simple generative model for CWPs in which pages acquire more visitors as their quality improves, but new visitors also have less chance of being able to contribute new information to a page as the page's quality improves. We evaluate our model primarily by examining the actual rate of editing of pages from Wikipedia, but also by examining commenting data from posts on the Russian section of the LiveJournal blog portal. Our CWP model reproduces all the salient features of the edit dynamics in the wiki and blog data – in particular, our model captures both the observed rise in edit rate after a page is founded and the ultimate $(1/t)$ decay in the edit rate after hitting a peak.

2. A GENERATIVE MODEL FOR COLLECTIVE WISDOM PROCESSES

An edit in a CWP is the result of someone adding meaningful information; it therefore requires that this visitor has information to add. By contrast, in models of pure arrival processes, an arrival always entails the addition of something new (e.g. a new link appearing in the Web graph). In general, meaningful edits improve the state of a CWP – the more edits a CWP receives, the higher its quality and the more credible it becomes. Since a better CWP is likely to attract more visitors, the more credible a page becomes, the more visible it becomes, attracting users at a faster rate. All else being equal, the higher the arrival rate, the more likely it is that someone will come along who has something to contribute to the page. Every user is endowed with some subset of the information on the topic of a CWP. There is some fixed, bounded total amount of information which is available, and so as a CWP improves, it is less likely that a new user's information set will contain anything new. We summarize these two interacting effects in the following observations.

OBSERVATION 1 (RICH GET RICHER). *An edit improves a CWP, increasing the visibility and hence the arrival rate of users.*

OBSERVATION 2 (INFORMATIONAL LIMIT). *The total available information of a CWP is bounded, so an improved CWP is less likely to be edited.*

2.1. A General Model

In order to formalize these observations, assume that a CWP is born at time 0. Let $t = 0, 1, \ldots$ denote the time step after birth. The state of the CWP at time t is represented by its information value $I_t \geq 0$ and its visibility $V_t \geq 0$. At time t, a user may arrive, carrying information value $X_t \geq 0$ drawn from some distribution, independently of the information brought by any previous users. If $X_t > I_t$, the user has more information than is already in the CWP and the user improves the CWP. In theory, I_t and X_t are sets of information, but without much loss in generality, we can represent them as real numbers. I_t and V_t are the state random variables in a stochastic dynamical system driven by the random variable X_t.

Intuitively, past visibility determines the probability of future arrivals. Visibility at a previous time step depends on the information value (credibility of the CWP). If a user arrives, she may improve the quality and hence affect the visibility. Let ρ_t be the probability that a user arrives at time t. We model ρ_t as a function of a base arrival probability ρ_0 and a visibility effect. Formally, $\rho_t = \rho_0 + \lambda V_{t-1}$, where $\lambda \in [0, 1 - \rho_0]$ is a parameter and $V_t \in [0, 1]$. This can capture in a simple manner processes with different base arrival rates and different multipliers for how the visibility of that process affects the arrival rate of users. The model thus provides some flexibility for different processes, while at the same time it is relatively easy to find linear fits for

particular processes. Of course one can generalize to more complex arrival processes, but the linear model is already quite powerful.

With probability $1 - \rho_t$, a user does not arrive at time t and, effectively, $X_t = 0$. Otherwise, with probability ρ_t, a user arrives, bringing information value $X_t > 0$. For $\lambda > 0$, the random variable X_t depends on V_{t-1} and there is an indirect dependence of X_t on I_{t-1}. A plausible information update rule is that an arriving user adds some fraction α of the value she could possibly add to a CWP. In this case, if $X_t > I_{t-1}$, then the value of the CWP gets augmented to $I_t \leftarrow (1-\alpha)I_{t-1} + \alpha X_t$.

2.2. A Simple Realization of the Model

First, consider the edit dynamics for the simplest realization of the above process, the *pure maximum process with no visibility*, for which $\lambda = 0$. In this case, $\forall t, \rho_t = \rho_0$, and $\alpha = 1$, so $I_t = \max_{\tau \le t} X_\tau$. We quantify the edit dynamics through the probability of an edit occurring at time t, $q_t = \Pr[\text{edit occurs at time } t] = \Pr[X_t > I_{t-1}]$.

THEOREM 2.1. *For the pure maximum process with no visibility, the probability of an edit at time t decays asymptotically at a $1/t$ rate.*

Proof:

$$q_t = \Pr[\text{edit at time } t] = \Pr[a_t X_t > \max\{X_0, a_1 X_1, \ldots, a_{t-1} X_{t-1}\}]$$

where a_t is an indicator variable indicating whether or not a user arrived at time t. Now, $\Pr[a_t X_t > \max\{X_0, a_1 X_1, \ldots, a_{t-1} X_{t-1}\}]$ is given by

$$\begin{aligned} & \rho_0 \int_{X_0}^{1} dF_X(x)\ \Pr[a_1 X_1 \le x; \ldots; a_{t-1} X_{t-1} \le x] \\ = \ & \rho_0 \int_{X_0}^{1} dF_X(x)(1 - \rho_0 + \rho_0 F_X(x))^{t-1} \\ = \ & \frac{1 - [1 - \rho_0(1 - X_0)]^t}{t} \end{aligned}$$

This completes the proof, because the exponentially decaying term is asymptotically negligible. □

All that is required in this proof is that X be a measurable random variable with probability measure dF_X, and the integral is defined in the Lebesgue sense. Note that X_0 is the information value at time 0, typically equal to 0.

While this theorem is only directly applicable to pure maximum processes with no visibility, the tail dynamics of typical CWPs will occur when the visibility has saturated. Therefore the asymptotic $1/t$ decay will carry over to general CWPs, and this can be seen in the asymptotic $1/t$ decay in edit rate in Wikipedia and the LiveJournal blogosphere (see Figure 2). We should also point out that the result applies for any choice of distribution from which the random variable X_t is drawn, and further, it does not even require the CWP to be bounded – i.e. the distribution of X_t can have unbounded support.

2.3. A General Solution

The pure maximum process with no visibility captures the effect of Observation 2 about CWPs, the informational limit. In doing so, it implies a continually decreasing edit rate (in fact, the edit rate even with $\alpha < 1$ would continually decrease). In contrast, CWPs in the real world tend to display a mid-life peak in edit rate. Incorporating a non-zero visibility effect ($\lambda > 0$) in the model yields exactly this behavior.

The arrival probability at time t is $\rho_t = \rho_0 + \lambda V_{t-1}$. We allow for some lag in the time it takes for a CWP's visibility to catch up to its quality, so $V_t = I_{t-\ell}$ (for simplicity, we assume a linear relationship). Blog posts may quickly be publicized to the readership of the blog through RSS feeds, for example, implying a small lag ($\ell \approx 0$). Wikipedia pages are largely accessed through search engines, so a newly improved Wikipedia page may only start experiencing increased traffic after a longer period related to the frequency with which search engines index the page ($\ell \approx 1$ month). We assume that $X_t \in [0,1]$ for concreteness.[1] Further, α need not be 1: we refer to the general process with $\lambda \in [0, 1-\rho_0]$ and $\alpha \in [0,1]$ as an *incremental CWP with lag*, which can be summarized by the following stochastic dynamical system:

$$\begin{aligned} V_t &= I_{t-\ell}, \\ \rho_t &= \rho_0 + \lambda V_{t-1}, \\ a_t &= \begin{cases} 0 & \text{w.p. } 1-\rho_t, \\ 1 & \text{w.p. } \rho_t. \end{cases}, \\ X_t &\sim F_X, \\ I_t &= \max\{I_{t-1}, (1-\alpha)I_{t-1} + \alpha X_t \cdot a_t\}. \end{aligned}$$

The initial conditions for the system are $I_t = 0$ for $t \leq 0$. The model is governed by the parameters $\lambda, \alpha, \rho_0, \ell$ and the distribution F_X from which X_t is drawn independently at each time step. The indicator variable a_t enforces $X_t = 0$ if no user arrives. The subtle dependency introduced by the visibility makes this apparently simple dynamical system quite challenging to solve. We can formulate an analytic solution which may be numerically solved through dynamic programming in a multi-dimensional function space, where the dimension is $\ell + 1$. For lag $\ell = 0$ this is a 1 dimensional dynamic program on a function space, which can be solved efficiently. For higher lag, the computational complexity of computing an accurate solution increases exponentially, and Monte Carlo simulation becomes the only realistic way to compute q_t.

Here we sketch the derivation of q_t for the special case of $\ell = 0$, and illustrate the edit dynamics that result from this model. For simplicity of exposition, we assume that information values are distributed uniformly on $[0,1]$. Let P_t be the distribution function for the information value I_t, $P_t(x) = \Pr[I_t \leq x]$. The edit probability

$$q_t = \Pr(I_t > I_{t-1}) = \int dx \; P_{t-1}(x) \Pr[I_t > x | I_{t-1} = x]$$

Integrating,

$$q_t = \int_0^1 dx \; P_{t-1}(x) \left(f(x)(\rho_0 + \lambda x) - \lambda(1 - F(x)) \right) = \int_0^1 dx \; P_{t-1}(x)(\rho_0 + 2\lambda x - \lambda)$$

where the second equality follows for the uniform distribution, and $f(x) = F'(x)$. We need to compute $P_t(x)$. Using the law of total probability,

$$P_t(x) = \rho_t \Pr(I_t \leq x | a_t = 1) + (1 - \rho_t) \Pr(I_t \leq x | a_t = 0).$$

Since $I_t \leq x$ if and only if $I_{t-1} \leq x$ and $(1-\alpha)I_{t-1} + \alpha a_t X_t \leq x$, we can relate $\Pr[I_t | a_t]$ to quantities involving the distribution of I_{t-1}, which is P_{t-1}. After some manipulation,

[1] All that is required is that the X_t are drawn from an integrable distribution.

we get:

$$P_t(x) = \begin{cases} Q_t(x) + (1-\rho_0-\lambda x)P_{t-1}(x) + \lambda G_{t-1}(x) & x \le \alpha, \\ Q_t(x) - Q_t(z) + zP_{t-1}(z)(\rho_0+\lambda z) - \lambda z G_{t-1}(z) + \\ \quad (1-\rho_0-\lambda x)P_{t-1}(x) + \lambda G_{t-1}(x) & x > \alpha. \end{cases}$$

where $z = \frac{x-\alpha}{1-\alpha}$,

$$Q_t(x) = xP_{t-1}(x)(\rho_0+\lambda x) - \left(\frac{\lambda x - (1-\alpha)\rho_0}{\alpha}\right) G_{t-1}(x) + 2\lambda\left(\frac{1-\alpha}{\alpha}\right) H_{t-1}(x)$$

and G_t, H_t are functions defined in terms of P_t:

$$G_t(x) = \int_0^x dy\ P_t(y), \qquad H_t(x) = \int_0^x dy\ yP_t(y)$$

Note that in this notation,

$$q_t = (\rho_0 - \lambda)G_{t-1}(1) + 2\lambda H_{t-1}(1)$$

2.4. Implications of the Model

Solving this model yields some important observations, which should reflect stylized facts in the data. The editing rate in any CWP follows a well-defined lifecycle: (1) it initially drops, up to a time equal to the lag; (2) at this point rising visibility takes over, and the edit rate reaches a peak; (3) finally, after the peak, when most of the information has been incorporated into the CWP, editing decays at an asymptotic $1/t$ rate.

3. EDIT DYNAMICS IN WIKIPEDIA

In order to verify our model, we analyze editing data for Wikipedia from its inception through May 24, 2008. While we are interested in information growth, not editing *per se*, it is impossible to directly measure growth in information; I_t (as defined above) is a hidden variable. We can only measure an edit, which would indicate (albeit not conclusively) that $I_{t+1} > I_t$. In order to make this correspondence as strong as possible with real-world data, we consider only *meaningful edits*, excluding edits attributed to vandalism or reversions of vandalism, and edits made by bots. Also, since the underlying arrival rate to Wikipedia as a whole has been increasing over time, while our generative model assumes a constant arrival rate, we normalize out the effect of the overall popularity of Wikipedia by adjusting the number of edits in a given day by the popularity of Wikipedia (as measured by Alexa's measurement of reach) on that day.

We focus only on pages that have received a significant number of edits (> 500, there are 43,616 such pages in our data) for several reasons. This paper is about the dynamics of collective information accrual: it is difficult to reach meaningful conclusions about the dynamics of editing or posting on wikis that have not received a sufficient number of edits. Further, such instances may be more indicative of individual opinion than of collective wisdom. Our sample selection allows us to focus on wikis that are indicative of collective processes at work. But, in the process is it possible that we ignore potentially important content? Actually, we find that pages that have received a large number of edits are disproportionately "important." There are two pieces of evidence for this.

First, Wilkinson and Huberman [2007] find that pages that are "featured" on Wikipedia (a proxy for quality) tend to have been edited a large number of times . This does not necessarily mean that high quality and high visibility articles *all* have

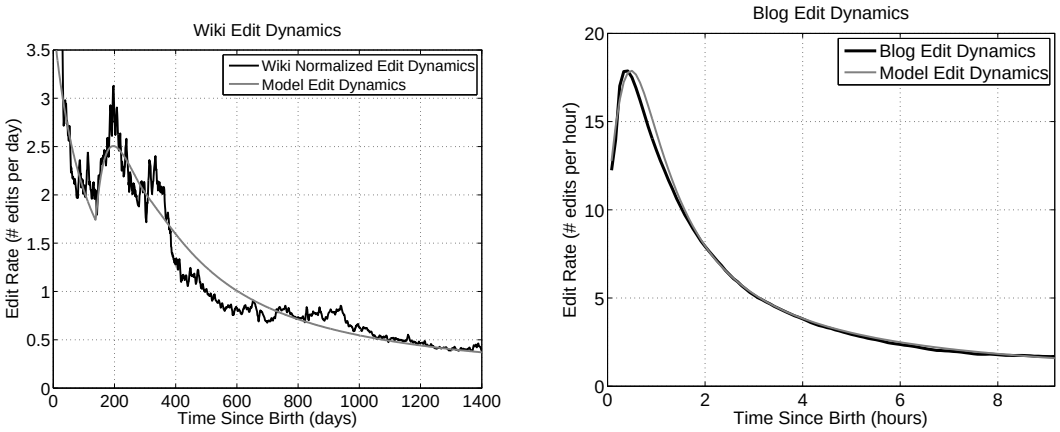

Fig. 1. (a) Wikipedia edit dynamics: average number of edits per day for heavily edited (> 500 edits) wikis, normalized by the popularity of Wikipedia. The model fit is using the CWP parameters $\lambda = 0.4, \alpha = 0.0045, \rho_0 = 0.14$, with one model time step being about 53 minutes. The Kolmogorov-Smirnov (L_∞) distance between the empirical CDF and model CDF is 0.0463 and the KL-divergence is 0.0858. (b) Blog edit dynamics: Average number of edits per 5 minute interval for heavily edited blog posts (> 50 comments). The model fit is using the CWP parameters $\lambda = 0.7, \alpha = 0.14, \rho_0 = 0.18$, with one model time step being about 52 seconds. The KS distance between the empirical CDF and model CDF is 0.0048 and the KL-divergence is 0.0010.

to be highly edited, so we conduct an empirical test, which provides the second piece of evidence for our hypothesis.

We examined a database collected by Spoerri [2007] of the 100 most popular pages on Wikipedia for five contiguous months from September 2006 to January 2007. This gives us 500 separate datapoints (230 unique pages). We checked the pages listed by Spoerri (or the pages they redirected to when searched on Wikipedia) and found that of these 500, 498 (228 unique pages) received more than 500 edits and were thus in our dataset. The two pages that did not make it into our dataset were clearly topics that received significant but brief media attention at the time. Additionally, only 5 other pages had less than 1000 edits, having between 500 and 1000, and each of these also only appeared on the monthly lists once. Therefore, 493 of the 500 data points had received more than 1000 edits as of May 2008. This indicates that a huge fraction of the most popular pages are also heavily edited.

Figure 1(a) shows the edit dynamics of the Wikipedia data and the best fit achieved from our model (using Monte Carlo simulation to minimize L_2 distance), demonstrating clearly the three regions of edit dynamics predicted by the model (1) an initial decay in the edit rate, followed by (2) a rise in the edit rate to a local maximum, followed by (3) decay to zero (in fact, at a $1/t$ rate – see Figure 2). The figure also reports some statistics about the goodness of fit. However, we provide these only as indicators: they cannot be used directly for statistical tests because the empirical distribution is an aggregate, and each of the individual distributions that makes up the aggregate is a highly dependent stochastic process: thus the usual assumptions about independent observations from the empirical CDF do not apply.

3.1. Another Dataset: Blog Dynamics

We further verify our model by fitting it to a blog dataset. While blogs are a different kind of social media from wikis, with significant differences in time scale and atten-

tional effects, highly commented blog posts are often also sources of information and opinion about a topic. We gathered all blog posts that received more than 50 comments (numbering 97,380) from the Russian segment of the LiveJournal blog provider from January to June of 2008.

Figure 1(b) shows the edit dynamics (number of comments received in consecutive 5-minute intervals since birth, averaged across all posts) for this data. In this case our model can be solved analytically for zero-lag, and the best fit computed. We see again that there is an excellent fit between model predictions and real data. The similarities in the edit rate dynamics for wikis and blogs are striking for the second and third stylized facts predicted by our model. The concave growth to a peak is in accordance with our generative model, and the asymptotic decay (documented at a finer level in Figure 2) closely matches the theoretical $1/t$ rate (we note that while Götz et al. [2009] find an exponent of -1.5 in the blogosphere, this is for post inter-arrival times rather than overall arrival times of comments on a particular post).

The major qualitative difference in this case is that the first stylized fact above, an initial "peak" at birth followed by immediate decay is present in the Wikipedia data but not in the blog data. This effect is captured in our generative model through the "no lag" assumption for blog data. This corresponds to a real-world difference in the nature of Wikipedia pages and blog posts: the visibility of blog posts is significantly less lagged than the visibility of Wikipedia pages, consistent with the hypothesis that new blog posts gain most of their visibility through regular readership and RSS feeds, while Wikipedia pages gain most of their visibility through delayed search-engine results (search engines are the primary source of traffic to Wikipedia, according to Nielsen[2]). Many of the pages we look at date to the early days of Wikipedia, when the popularity of Wikipedia pages was not high (early on in the life of the page). These pages had to become "trusted" (for example, highly linked-to) in order to rise in search engine rankings. In this context, it makes sense that Wikipedia pages would suffer a significant visibility lag.

An interesting quantitative difference is that the increment parameter α is much smaller for the best fit to the Wikipedia data than it is for the best fit to the blog data, indicative of the type of CWP: since blogs are conversational, a visitor is likely to contribute more of their opinion at one sitting, whereas wikis are archival and so require more detailed editing: hence users contribute a lower fraction of what they may theoretically be able to.

4. CONCLUSIONS

We have introduced a new, generative model for Collective Wisdom Processes, where users arrive asynchronously both looking for information and potentially able to contribute information. Our model involves the interplay of two key elements: (1) a rich-get-richer phenomenon, in which page quality improves with more edits, and higher quality pages attract more visitors who may be able to contribute information; and (2) an informational limit on growth, whereby new visitors are less likely to have something new to contribute to pages that are already high quality. When coupled with the possibility of a visibility "lag", this model captures the editing dynamics of observable CWPs.

Our stochastic model predicts three major stylized facts about the dynamics of CWPs. After creation, the edit rate may decrease up to a "visibility lag" time; it will then increase to a peak with concave growth; ultimately the edit rate will decrease at a $1/t$ rate. We find that the model is capable of fitting two different CWP datasets (highly edited Wikipedia articles and highly commented posts from the Russian blo-

[2]http://www.nielsen-netratings.com/pr/pr_080514.pdf

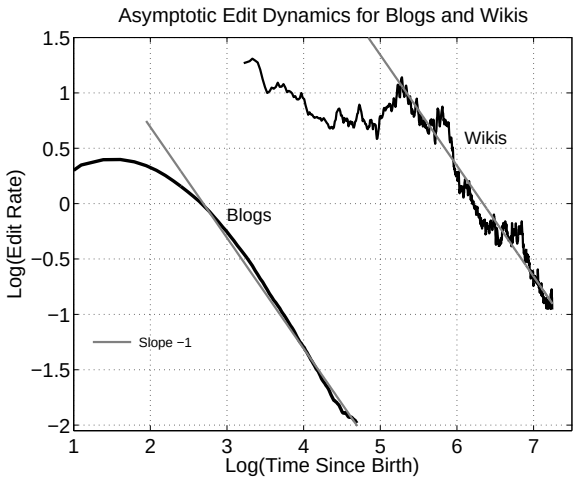

Fig. 2. The asymptotic decay in edit rate. The thin straight lines with slope -1 correspond to $1/t$ decay. The optimal linear fits to the tail of the blog data and Wikipedia data had slopes of -1.03 and -0.98 respectively. The tail for the blogs was the edit dynamics from about 1 hour to about 9 hours after birth, and for the wikis it was the edit dynamics from 450 days to 1400 days from birth.

gosphere), demonstrating its power. Our model is parsimonious, having few free parameters; therefore, it is unlikely to overfit and hence quite generalizable. We believe it would be difficult to attain equivalent predictive and explanatory power by fitting a standard parametric model. To provide one example of the usefulness of this generalizability, we conducted a simple prediction test: we fit the model to only the early stage of the edit distribution for blogs and predicted the number of total comments an average blog post would get. The results are promising: by fitting the model to the first hour of comments received, we could predict the total number of comments with 89.3% accuracy, and by fitting it to the first hour-and-a-half we could predict the total number of comments with 94.4% accuracy.

ACKNOWLEDGMENTS

We thank Konstantin Mertsalov for collecting and supplying us with the LiveJournal data.

REFERENCES

BARABASI, A. L. AND ALBERT, R. 1999. Emergence of scaling in random networks. *Science 286,* 5439, 509–512.

BUTLER, B., JOYCE, E., AND PIKE, J. 2008. Don't look now, but we've created a bureaucracy: The nature and roles of policies and rules in Wikipedia. In *Proc. SIGCHI*. ACM, 1101–1110.

CAPOCCI, A., SERVEDIO, V. D. P., COLAIORI, F., BURIOL, L. S., DONATO, D., LEONARDI, S., AND CALDARELLI, G. 2006. Preferential attachment in the growth of social networks: The internet encyclopedia Wikipedia. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics) 74,* 3.

DAS, S. AND MAGDON-ISMAIL, M. 2010. Collective wisdom: Information growth in wikis and blogs. In *Proceedings of the ACM Conference on Electronic Commerce*. 231–240.

FALOUTSOS, M., FALOUTSOS, P., AND FALOUTSOS, C. 1999. On power-law relationships of the internet topology. In *Proc. SIGCOMM*. Vol. 29. ACM Press, New York, NY, USA, 251–262.

GILES, J. 2005. Internet encyclopaedias go head to head. *Nature 438,* 7070, 900–901.

GÖTZ, M., LESKOVEC, J., MCGLOHON, M., AND FALOUTSOS, C. 2009. Modeling blog dynamics. In *International Conference on Weblogs and Social Media*. 26–33.

KITTUR, A., SUH, B., PENDLETON, B. A., AND CHI, E. H. 2007. He says, she says: Conflict and coordination in Wikipedia. In *Proc. SIGCHI*. ACM Press, New York, NY, USA, 453–462.

KUMAR, R., NOVAK, J., RAGHAVAN, P., AND TOMKINS, A. 2004. Structure and evolution of blogspace. *Commun. ACM 47,* 12, 35–39.

LESKOVEC, J., BACKSTROM, L., KUMAR, R., AND TOMKINS, A. 2008. Microscopic Evolution of Social Networks. In *Proceedings of the Conference on Knowledge Discovery and Data Mining*. ACM Press, 462–70.

SILVA, L., GOEL, L., AND MOUSAVIDIN, E. 2009. Exploring the dynamics of blog communities: The case of MetaFilter. *Information Systems Journal 19,* 1, 55–81.

SPINELLIS, D. AND LOURIDAS, P. 2008. The collaborative organization of knowledge. *Commun. ACM 51,* 8, 68–73.

SPOERRI, A. 2007. What is popular on Wikipedia and why? *First Monday 12,* 4.

VAZ DE MELO, P., AKOGLU, L., FALOUTSOS, C., AND LOUREIRO, A. 2010. Surprising patterns for the call duration distribution of mobile phone users. In *Proc. PKDD*. 354–369.

WILKINSON, D. M. AND HUBERMAN, B. A. 2007. Assessing the value of coooperation in Wikipedia. *First Monday 12,* 4.

WU, F. AND HUBERMAN, B. 2007. Novelty and collective attention. *Proceedings of the National Academy of Sciences 104,* 45, 17599.

Received June 2009; revised June 2011; accepted July 2011