Collective Wisdom: Information Growth in Wikis and Blogs

Sanmay Das
Dept. of Computer Science
Rensselaer Polytechnic Institute
Troy, NY 12180
sanmay@cs.rpi.edu

Malik Magdon-Ismail
Dept. of Computer Science
Rensselaer Polytechnic Institute
Troy, NY 12180
magdon@cs.rpi.edu

ABSTRACT

Wikis and blogs have become enormously successful media for collaborative information creation. Articles and posts 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 and on blogs collected from LiveJournal; in particular, it captures the observed rise in the edit rate, followed by $1/t$ decay.

Categories and Subject Descriptors

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

General Terms

Algorithms, Human Factors, Measurement, Theory

Keywords

Collective Intelligence; Social Networks

1. INTRODUCTION

Wikis and blogs have become trusted sources of information for most web users. People use blogs as sources of news and opinion and they turn to Wikipedia for information on specific topics. Independent studies have verified that Wikipedia articles are of comparable quality to the Encyclopedia Britannica [6]. The success of these media for

collaborative information creation and dissemination leads to new questions about the dynamic processes that create trusted sources of information [13].

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. In this paper we examine the dynamics of editing in CWPs, specifically using data from Wikipedia and from Russian users of the Live-Journal blog portal. Scientific studies of Wikipedia and the Blogosphere as social systems have focused on growth (addition of new articles) [4, 13], on macroscopic properties of communication dynamics [12], and on network modeling of the implied social interactions [9]. Recently there has also been some focus on what makes content popular [14, 16]. In terms of modeling, this work generally looks at decaying popularity directly as a function of decaying novelty; for example, Wu and Huberman [16, 17] develop a model of popularity where a multiplicative, exponentially decaying novelty factor attenuates "rich-get-richer" phenomena. Novelty is treated as an exogenous variable in these models, so the decay of attention is imposed on the model. Leskovec et al. [11] recently studied the dynamics of the news cycle using a new algorithm for finding "memes" in news media and blogs. They also propose a model that couples imitative dynamics among media outlets with a decaying recency factor.

Another direction of research has studied the sociological implications of these new media, examining the editing behavior of users [8] and the emergence of bureaucracies in Wikipedia [3, 7]. While this paper is related to the entire literature on the growth of networks, both general and social [1, 10], it differs in its focus on the dynamics of information rather than on the dynamics of user arrivals and departures. Our model is similar in spirit to Bass' work on the timing of adoption of new products in a social system [2]. Bass models the adoption of products by innovators and imitators in a population; since his model is for durable goods, there is a natural market saturation phenomenon.

Wikis and blogs 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 and blogs. Wilkinson and Huberman [15] 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. Similarly, we study blog posts that receive a high number of comments.

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 log-normal 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 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 [1] or the Internet [5], 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. We present data on the actual rate of editing of (1) pages from Wikipedia, and (2) posts from the Russian section of the LiveJournal blog portal. The data confirms the hypothesis that the rate of editing decays after reaching a peak. We propose a simple 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. 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. EDIT DYNAMICS ON WIKIS AND BLOGS

We present and analyze two sets of data on the dynamics of editing behavior in CWPs. The first of these is editing data for Wikipedia from its inception through May 2008, and the second is comment posting data from the Russian segment of the LiveJournal blog provider from January to June of 2008. We consider Wikipedia pages with more than 500 edits, and all blog posts that received more than 50 comments. Additionally, we consider only the *meaningful edits* for Wikipedia pages. This definition excludes edits attributed to vandalism or reversions of vandalism, and edits made by bots. Details of the data and processing techniques are below.

2.1 Why Highly Edited Documents?

We focus only on pages that have received a significant number of edits 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 and blogs 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 and blogs that are undoubtedly indicative of collective processes at work, but in the process is it possible that we ignore potentially more 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 [15] 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 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 [14][1] 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 were not were clearly topics that received significant but brief media attention at the time, namely "Buggery Act 1533" and "Katie Rees." Additionally, only 5 other pages had less than 1000 edits, having between 500 and 1000 (these were "Tara Conner," "List of female porn stars by decade," "History of the board game Monopoly," "Robert Gates," and "Operation Ten-Go"), 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.

2.2 Wikipedia Data Collection

From the history of all edits to all Wikipedia pages as of May 24, 2008,[2] we extracted all pages with more than 500 total edits (there were 43,616 such pages). We removed edits that were performed by automated bots, and aggregated sequences of edits by a single user into a single edit. We also used a naïve heuristic to remove posts that could be attributed to vandalism, or reversions of vandalism – the heuristic looked for one of a set of keywords[3] and removed any edit in which one of the keywords occurred, as well as its predecessor edit. We refer to the remaining edits as *meaningful edits*.

2.3 Edit Dynamics

Wikipedia Dynamics.
Figure 1 shows the growth in number of edits per day for Wikipedia articles that received more than 500 edits in total, along with the growth in popularity of Wikipedia it-

---

[1]Available at `http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1765/1645`
[2]From `http://download.wikimedia.org/enwiki/20080524/`
[3]The keywords were: *neutral, pov, remove, replace, revert, rv, undid, vandal, view.*
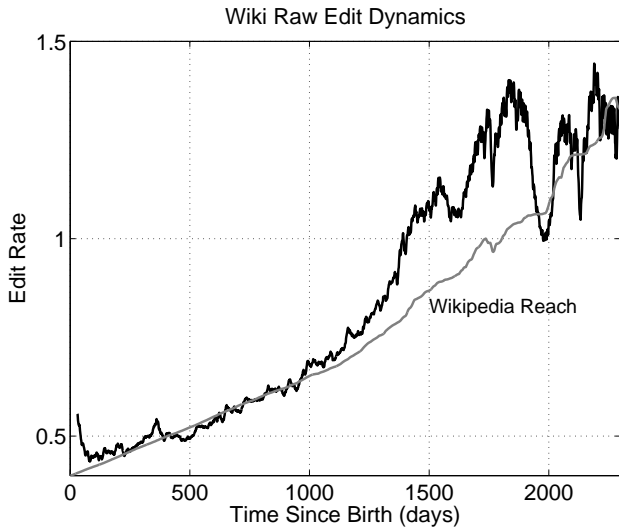
Figure 1: The number of meaningful edits per day averaged over the 43,616 wikis that received more than 500 edits. For each wiki, birth (time 0) is defined as the time of the first post. The graph also shows the (rescaled) popularity of Wikipedia, as measured by the Alexa reach factor, also averaged over the 43,616 pages from their birth times.

self. The edit rate initially grows at what appears to be an exponential rate, but then, despite the continuing increase in the popularity of Wikipedia, the edit rate starts to decay. This data falsifies any pure growth model, including "rich-get-richer" models.

Normalization.

Given the rate at which the popularity of Wikipedia as a whole has been growing, it is not even obvious whether the initial exponential growth phase should be attributed to this Wikipedia-wide growth, or to individual page effects. The edit dynamics curve shown in Figure 2(a) normalizes out the effect of the overall popularity of Wikipedia by adjusting the number of edits in a given day by the popularity of Wikipedia on that day (popularity is defined by Alexa's measurement of reach, details are in Appendix A).

Blog Dynamics.

Figure 2(b) shows the edit dynamics (number of comments received in consecutive 5-minute intervals since birth, averaged across all posts in the dataset) for data from the Russian LiveJournal Blogosphere. In this case normalizing is unnecessary because the popularity of LiveJournal is essentially constant in the first half of 2008, the timeframe over which the data was collected. LiveJournal provides a news feed which is updated whenever a new post is made. We monitored this feed over the first half of 2008 for posts with Cyrillic characters, indicative of a post by the Russian segment of LiveJournal. LiveJournal is the almost exclusive provider of blog resources to the Russian population, and the Russian segment of LiveJournal is particularly amenable to study – it is a self-contained sub-population with hundreds of

thousands of members. The total dataset included 4,874,567 blog posts. For each post, we obtained the entire sequence of comments as of two weeks from the initial posting date (in our experience two weeks is the maximum lifetime of a post except in a vanishingly small number of cases). We consider posts with more than 50 comments (there are 97,380 such posts).

2.4 Comparing Wikis and Blogs

The similarities in the edit rate dynamics for wikis and blogs are striking. For Wikipedia, the edit rate initially drops, then rises to a local peak after which it decays down toward zero. The dynamics are similar in the Blogosphere, except that the initial decay is not present. It is clear from the data that articles do not continue to accrue edits at an ever increasing rate, so pure growth models are not viable explanations. Further, the initial growth in edit rate does not appear to be exponential, at least for the blogs, ruling out even an initial phase of geometric growth. The data for Wikipedia is noisier, but qualitatively similar.

The difference in the appearance of the initial "peak" at birth in the Wikipedia data but not the blog data is interesting in several ways. Part of this difference can be explained by the nature of Wikipedia pages and blog posts. Wikipedia pages often start off as requests for information or for people to write more about a topic, which can lead to an initial burst of editing, whereas blog posts are often well thought out, fairly comprehensive posts on an issue (we remove some of this effect in the data by accumulating multiple sequential edits by the same author in Wikipedia into one "edit," but we are unable to capture, for example, repeated alternating edits by two people that are not just reversions of each others' edits). We will also see later that a substantial fraction of this effect can be explained by positing the existence of a lag or delay in the popularity of Wikipedia pages catching up to the quality of their content, which may have been a very significant factor, especially in the early days of Wikipedia, whereas blogs have always been more of an immediate medium which people read regularly.

2.5 The Dynamics of CWPs

Thus, the data allows us to break the edit dynamics of a CWP into three regions: (1) an initial decay in the edit rate (which may not be present), 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 in both cases – see Figure 3). Wikis and blogs are CWPs of distinctly different natures, in terms of both content and the time-scale of editing dynamics. The data indicates that the above stylized facts may be universal to CWPs, yielding a picture of the dynamics for a general CWP: after creation of an article, a topic, or a blog post, the edit rate may decrease; it will then increase to a peak, seemingly with concave growth; ultimately the edit rate will decrease at a $1/t$ rate. Any model of a CWP must be consistent with these facts. In the next section we propose a simple model based on two observations about CWPs – first, higher quality articles attract more attention and more visitors, and second, the limit on information available on a topic limits editing behavior. We show how the interaction of these two effects is sufficient to produce the exact phenomena observed in the data, and demonstrate the fit of the model to the data. We also discuss how the model can then be used to make nontrivial inferences about unobserved
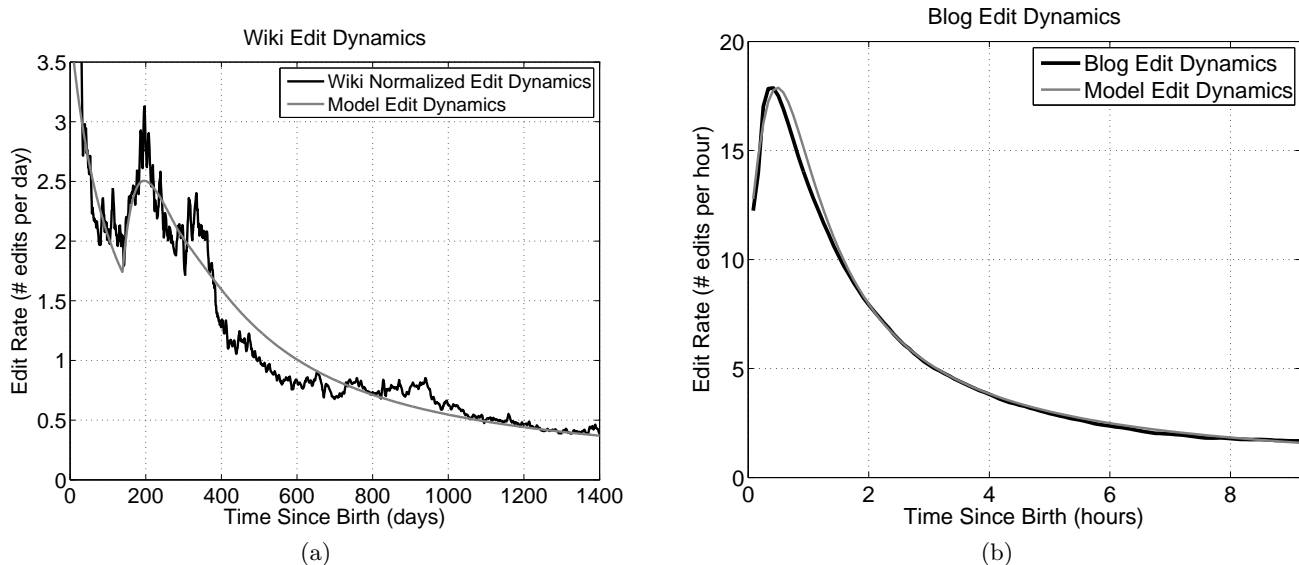
Figure 2: (a) Wikipedia edit dynamics: average number of edits per day for heavily edited (> 500 edits) wikis, normalized by the popularity of Wikipedia. The model fit is using the CWP parameters $\lambda = 0.4, \alpha = 0.0045, \rho_0 = 0.14$, with one model time step being about 53 minutes. (b) Blog edit dynamics: Average number of edits per 5 minute interval for heavily edited blog posts (> 50 comments, 97,380 such blogs). 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.

variables, like overall traffic to Wikipedia, or the amount of their knowledge that users contribute in different media.

3. 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. In contrast, a visitor in a pure arrival process (say the WWW) will always add something (a new link). 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 (for example, [WH] show that heavily edited Wikipedia pages tend to be of higher quality). 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.*

3.1 A General Model

In order to formalize these observations, assume that a CWP is born at time 0. Let $t = 0, 1, \dots$ 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 $I_t \leftarrow (1-\alpha)I_{t-1} + \alpha X_t$.

3.2 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 $I_t = \max_{\tau \leq t} X_\tau$. 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 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

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

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 3). 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.

3.3 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[4] , so $V_t = I_{t-\ell}$ (for simplicity, we assume a linear relationship). Intuitively, this means that the visibility of a page is a function of its quality ℓ steps ago. 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 – for example, the search engine ranking of a page may be proportional to how good the page was several weeks or months in the past, when people started linking to it, rather than its quality right now, since the links may not yet have appeared because people have not seen how much it improved in the last few days, and are still not willing to link to it as a source. We assume that $X_t \in [0,1]$ for concreteness.[5] 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:

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

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 $a_t X_t = 0$ 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.[6]

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 X_t are distributed uniformly on $[0,1]$. The detailed derivation can be found in Appendix B.

Let P_t be the distribution function for the information value I_t, $P_t(x) = \Pr[I_t \leq x]$. The edit probability

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

---

[4] Of course, visibility is always a fews steps behind quality, but since the rate of growth of quality will slow down, the visibility will eventually become very close to the current quality.

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

[6] Using the dynamic programming approach, the complexity of computing q_t out to time step T with accuracy $O(\epsilon)$ increases according to $\Omega(\frac{T}{\epsilon^{\ell+1}})$.

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 last line 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, we get the following equation for $P_t(x)$:

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

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

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

G_t, H_t are functions defined in terms of P_t:

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

Note that in this notation,

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

3.4 Implications of the Model

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

Figure 2(b) shows that the model with zero lag closely replicates the observed edit dynamics of the Blogosphere, and Figure 2(a) shows the model fit to the Wikipedia data with non-zero lag (the Wikipedia results are computed using Monte Carlo simulation). The asymptotic decay is documented at a finer level in Figure 3, which shows that the observed data closely matches the theoretical $1/t$ rate.

The above results demonstrate that our model provides a good fit for Wikipedia pages that received more than 500 edits and blog posts that received more than 50 comments. How can we validate the predictive ability of the model? One test is to see if fitting the model on the initial dynamics can successfully predict future dynamics. We fit the model to the initial history (1, 1.5, and 2 hours) of aggregate blog data. We then predict the number of future edits the blog will receive through approximately the first 9 hours of its life. Figure 4 shows the predicted curve based on 1.5 hours of editing history, showing that the model is very good at
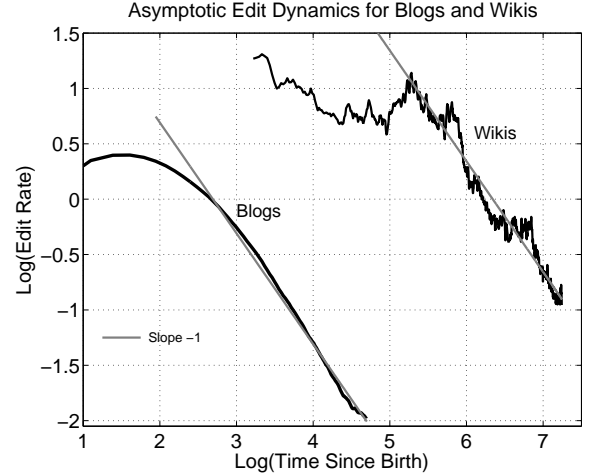

Figure 3: The asymptotic decay in edit rate. The thin straight lines with slope -1 correspond to $1/t$ decay. The optimal linear fits to the tail of the blog data and Wikipedia data had slopes of -1.03 and -0.98 respectively. The tail for the blogs was the edit dynamics from about 1 hour to about 9 hours after birth, and for the wikis it was the edit dynamics from 450 days to 1400 days from birth.

predicting the tail even using a short history of initial dynamics. The actual number of future edits predicted for different training lengths is:

Training data	Predicted # comments	Real # comments	Prediction accuracy
1 hour	29.2	32.7	89.3%
1.5 hours	25.5	27.0	94.4%
2 hours	21.4	22.6	94.7%

3.5 More complex models

The model proposed here is too simple to be an accurate depiction of how information update processes that involve humans actually work. In particular, it may be more reasonable to imagine each person having an information set containing many different elements of information, so that anyone with a new piece of information could add something to a page. Further, different people may have different types of information that they can contribute to a page – for example, someone with a lot of information about a topic may have poor grammar and spelling, so that someone who views the page after them, but knows little about the topic, may be able to substantially improve the page (which they could not have done if the previous person had not added their knowledge) – this would violate the independence assumption we make. However, the model presented here is simple enough to remain analytically tractable in many cases, while at the same time it is demonstrably good at explaining all the major stylized facts in the data. Monte Carlo experiments with more complex models, including the information set model mentioned above, do not provide any substantially new insights into CWPs that are not already in this simpler model. In particular, minor aspects of the edit dynamics may change but the overall shape (including the $1/t$ decay)
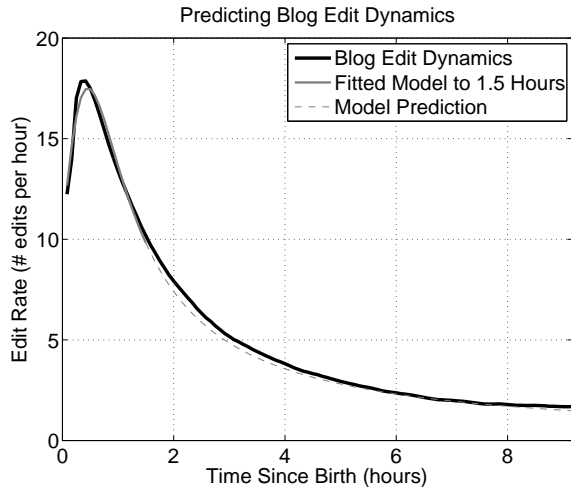
Figure 4: Results of fitting the model to the first 1.5 hours of blog data and then predicting the next 8 hours (approximately). The model predicts that 26 more comments will be made in the next 8 hours, while in reality about 27 comments are actually made. Fitting the model to 1 hour of data leads to a little more underestimation and fitting the model to 2 hours of data leads to a slightly better fit of the tail.

does not. There may be some domain-specific benefits to finer models, but the appeal of this model is its simplicity and applicability to diverse CWPs like wikis and blogs.

4.   DISCUSSION

Wikipedia and the Blogosphere both provide venues for the creation and aggregation of knowledge. At the same time, these are very different types of CWPs. Wikipedia is mainly a medium for the creation of archival articles. Blogs are a conversational medium, in which users contribute their views on topics in a more dynamic manner. Despite the differences in the scale and dynamics of these media, they exhibit some remarkable regularities. The data demonstrates that articles have a natural life cycle, defined by (possibly delayed) initial growth in edit rate, followed by $(1/t)$ decay. This data falsifies any attempt to explain CWPs using pure growth models analogous to those used to explain the continued growth of networks.

What properties of CWPs lead to this apparently universal behavior? We propose a model involving 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.

The model also makes it possible to make nontrivial inferences by fitting real data as in Figure 2. In the previous section we have presented one example, in which we show that we can predict future edit dynamics successfully given

prior edit dynamics. We present some more examples in this section. First, we can infer that 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[7]). Many of the pages we look at date to the early days of Wikipedia, when the popularity of Wikipedia pages was not high from 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.

Second, the model we propose in this paper suggests that significant editing leads to high quality, and high quality in turn to high visibility (the mechanism could be through improvement in search engine rankings as mentioned above, for example). Wilkinson and Huberman [15] have previously demonstrated a correlation between being highly edited and being high quality . Our model implies that the most popular pages on Wikipedia should be relatively highly edited ones. Indeed, an analysis of Spoerri's (2007) data (the 100 most popular Wikipedia pages for 5 contiguous months from Sep 2006-Jan 2007) shows that of these 500 observations (230 unique pages), 498 (228 unique pages) received more than 500 edits, and 493 (223 unique pages) received more than 1000 edits (see Supplementary Information for further details).

The final example 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.

This paper focuses on the edit dynamics of CWPs. For wikis and blogs we measure the typical edit dynamics by averaging across articles. It would also be interesting to consider the full distribution of edit rates (as opposed to the average) at a particular time step. This cross sectional distribution appears to be non-Gaussian in the tail (see also [15]). Our model addresses the edit dynamics of a *single* CWP. To understand the cross sectional edit rate distribution of a collection of CWPs, it would be necessary to additionally model the distribution from which the parameters of each CWP are drawn, paying particular attention to correlations between the parameters. For example, there could be a strong negative correlation between the initial popularity of a CWP and the increment parameter α. These correlations, together with distributions over the birth times of different articles could be fit to the overall distribution of edits to articles reported in other studies [4, 15].

5.   ACKNOWLEDGMENTS

---

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

References

[1] A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, October 1999.

[2] F.M. Bass. A new product growth model for consumer durables. *Management Science*, 15(5):215–227, 1969.

[3] Brian Butler, Elisabeth Joyce, and Jacqueline Pike. Don't look now, but we've created a bureaucracy: The nature and roles of policies and rules in Wikipedia. In *CHI '08: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1101–1110, New York, NY, USA, 2008. ACM.

[4] A. Capocci, V. D. P. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli. Preferential attachment in the growth of social networks: The Internet encyclopedia Wikipedia. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 74 (3), 2006.

[5] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationships of the Internet topology. In *SIGCOMM '99: Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, volume 29, pages 251–262, New York, NY, USA, October 1999. ACM Press.

[6] Jim Giles. Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901, December 2005.

[7] A. Kittur, E. H. Chi, B. A. Pendleton, B. Suh, and T. Mytkowicz. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *CHI '07: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2007.

[8] Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. He says, she says: Conflict and coordination in Wikipedia. In *CHI '07: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 453–462, New York, NY, USA, 2007. ACM Press.

[9] Ravi Kumar, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. Structure and evolution of blogspace. *Communications of the ACM*, 47(12):35–39, December 2004.

[10] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *KDD '08: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 462–70. ACM Press, 2008.

[11] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *KDD '09: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 497–506. ACM New York, NY, USA, 2009.

[12] Leiser Silva, Lakshmi Goel, and Elham Mousavidin. Exploring the dynamics of blog communities: the case of Metafilter. *Information Systems Journal*, 9999(9999), 2008.

[13] Diomidis Spinellis and Panagiotis Louridas. The collaborative organization of knowledge. *Communications of the ACM*, 51(8):68–73, August 2008.

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

[15] Dennis M. Wilkinson and Bernardo A. Huberman. Assessing the value of cooperation in Wikipedia. *First Monday*, 12(4), Feb 2007.

[16] F. Wu and B.A. Huberman. Novelty and collective attention. *Proceedings of the National Academy of Sciences*, 104(45):17599, 2007.

[17] F. Wu and B.A. Huberman. Popularity, novelty, and attention. *Proceedings of the ACM Conference on E-Commerce*, pages 240–245, 2008.

APPENDIX

A. WIKIPEDIA NORMALIZATION

The raw number of meaningful edits received for each page every day since birth was computed. We then divided this number by a measure of the reach of Wikipedia on that day, as measured by Alexa reach per million data downloaded from Wikipedia[8] There are two points to note about this normalization. (1) Since the data is in general either weekly or biweekly and noisy, we took the average of the reach at the previous and next measurement dates. (2) The Alexa data does not reach back to the time of birth of the first page in our sample. We use a conservative linear interpolation, assuming that the reach of Wikipedia was 1 at the first birth time in our sample, and interpolating to a reach of 125 (91 weeks later). Many of the pages in our sample did exist in this early time period in the history of Wikipedia. Note that the main effect of going down all the way to a reach of 1 is probably to slightly inflate the initial editing peak we see at birth – this is not central to any of the arguments. The data on the average normalizing factors over the most highly edited pages in our sample can be seen in Figure 1. The time series for all pages were aggregated by aligning the time series at time 0 for the birth of each page. This aggregated time series appears in Figure 2 (a).

B. DERIVATION OF EDIT DYNAMICS

In this section we use the same notation as in the main text. To recap, the information value of a CWP at time t is given by I_t, let $P_t(x) = \Pr[I_t \leq x]$. Let $p_t(x)$ be the density function corresponding to $P_t(x)$, $p_t(x)dx = \Pr[I_t \in (x, dx)]$.

---

[8] http://en.wikipedia.org/wiki/Wikipedia:Awareness_statistics

B.1 Derivation of q_t

This is the case where lag $l = 1$. We want the probability of an information update, $q_t = \Pr[I_t > I_{t-1}]$

$$= \int_0^1 dx \ \Pr[I_t > x|I_{t-1} \in (x, dx)] \Pr[I_{t-1} \in (x, dx)],$$

$$= \int_0^1 dx \ \Pr[I_t > x|I_{t-1} \in (x, dx)]p_{t-1}(x).$$

We have to be careful when manipulating this integral because there is a δ-function at $x = 0$ in $\Pr[I_{t-1} \in (x, dx)]$. Breaking up the integral into two parts, we get

$$P_{t-1}(0)\rho_0 + \int_0^1 dx \ (\rho_0 + \lambda x)(1 - F(x))p_{t-1}(x)$$

The contribution of the δ-function has been extracted explicitly in the first term, and $\Pr[I_t > x|I_{t-1} \in (x, dx)]$ has been written explicitly as the combination of the probability of user arrival at time t and that user having something to contribute. Since $p_{t-1}(x) = P'_{t-1}(x)$, we can perform integration by parts. The integral in the above expression then becomes:

$$\left[(\rho_0 + \lambda x)(1 - F(x))P_{t-1}(x)\right]_0^1 -$$
$$\int_0^1 dx \ P_{t-1}(x)[(\rho_0 + \lambda x)(-f(x)) + \lambda(1 - F(x))]$$

where $f(x) = F'(x)$. Simplifying further gives us

$$\int_0^1 dx \ P_{t-1}(x)[f(x)(\rho_0 + \lambda x) - \lambda(1 - F(x))] -$$
$$\rho_0(1 - F(0))P_{t-1}(0)$$

Assuming that $F(0) = 0$, the term involving $P_{t-1}(0)$ cancels with the term extracted above to give

$$q_t = \int_0^1 dx \ P_{t-1}(x)[f(x)(\rho_0 + \lambda x) - \lambda(1 - F(x))].$$

Once we have computed $P_t(x)$, a straightforward numerical integration gives q_t. For uniformly distributed X_t ($F_X(x) = x$, $f_{val}(x) = 1$),

$$q_t = \int_0^1 dx \ P_{t-1}(x)(\rho_0 + 2\lambda x - \lambda).$$

B.2 Derivation of P_t

We would like to compute the probability distribution for the information value I_t. In our notation P_t is the cumulative distribution function and p_t is the density function.

$$P_t(x) = \Pr[I_t \le x; a_t = 1] + \Pr[I_t \le x; a_t = 0],$$
$$= \Pr[I_t \le x|a_t = 1]\rho_t + \Pr[I_t \le x|a_t = 0](1 - \rho_t),$$
$$= \Pr[I_t \le x|a_t = 1]\rho_t + \Pr[I_{t-1} \le x|a_t = 0](1 - \rho_t).$$

We continue by evaluating $\Pr[I_t \le x|a_t = 1]$,

$$= \Pr[I_{t-1} \le x; (1 - \alpha)I_{t-1} + \alpha X_t \le x|a_t = 1],$$
$$= \int_0^x dy \ \Pr[I_{t-1} \in (y, dy)|a_t = 1]F_X(\tfrac{x - (1-\alpha)y}{\alpha}).$$

By Bayes theorem, $\Pr[I_{t-1} \in (y, dy)|a_t = 1]$ is given by

$$(\Pr[a_t = 1|I_{t-1} \in (y, dy)] \Pr[I_{t-1} \in (y, dy)]) / \Pr[a_t = 1]$$

We will now specialize[9] to the case when the lag $\ell = 0$ and F_X is the uniform distribution, in which case

$$\Pr[a_t = 1|I_{t-1} \in (y, dy)] = \Pr[a_t = 1|V_{t-1} \in (y, dy)]$$
$$= \rho_0 + \lambda y$$

In this case, $V_t = I_{t-1}$ and we have

$$\Pr[I_t \le x|a_t = 1] = \frac{1}{\rho_t} \int_0^x dy \ (\rho_0 + \lambda y)p_{t-1}(y)F_X(\tfrac{x - (1-\alpha)y}{\alpha})$$

For the uniform distribution when $x \ge 0$, $F_X(x) = \min(x, 1)$.

$$\Pr[I_t \le x|a_t = 1] = \frac{1}{\rho_t} \int_0^x dy \ (\rho_0 + \lambda y)p_{t-1}(y) \min(\tfrac{x - (1-\alpha)y}{\alpha}, 1)$$

There are two cases in this integral. The first is when $x \le \alpha$, in which case $\min(\tfrac{x - (1-\alpha)y}{\alpha}, 1) = \tfrac{x - (1-\alpha)y}{\alpha}$. Then the integral becomes

$$\frac{1}{\alpha\rho_t} \int_0^x dy \ (\rho_0 + \lambda y)p_{t-1}(y)(x - (1-\alpha)y)$$
$$= \frac{1}{\alpha\rho_t} \int_0^x dy \ p_{t-1}(y)(\rho_0 x + (\lambda x - \rho_0(1-\alpha)y)y - \lambda(1-\alpha)y^2)$$

which can be rewritten as

$$\frac{1}{\alpha\rho_t} \Big[\alpha x P_{t-1}(x)(\rho_0 + \lambda x) -$$
$$\int_0^x dy \ (\lambda x - (1-\alpha)(\rho_0 + 2\lambda y))P_{t-1}(y)\Big]$$

where $G_t(x) = \int_0^x dy \ P_t(y)$, and $H_t(x) = \int_0^x dy \ yP_t(y)$. For convenience, we introduce the function:

$$Q_t(x) = \int_0^x dy \ (\rho_0 + \lambda y)p_{t-1}(y) \left(\frac{x - (1-\alpha)y}{\alpha}\right)$$

Then, for $x \le \alpha$, $\Pr[I_t \le x|a_t = 1] = \frac{1}{\rho_t}Q_t(x)$.

When $x > \alpha$, we break up the integral into two parts. Let $z = \frac{x - \alpha}{1 - \alpha}$. Then the integral becomes

$$\frac{1}{\rho_t} \int_0^z dy \ (\rho_0 + \lambda y)p_{t-1}(y) +$$
$$\frac{1}{\alpha\rho_t} \int_z^x dy \ (\rho_0 + \lambda y)p_{t-1}(y)(x - (1-\alpha)y)$$

$$= \frac{1}{\rho_t}[Q_t(x) - Q_t(z) + zP_{t-1}(z)(\rho_0 + \lambda z) - \lambda zG_{t-1}(z)]$$

where the last line follows after some elementary manipulations. We use similar logic to manipulate $\Pr[I_{t-1} \le x|a_t = 0]$:

$$\Pr[I_{t-1} \le x|a_t = 0] = \int_0^x dy \ \Pr[I_{t-1} \in (y, dy)|a_t = 0]$$
$$= \frac{1}{1 - \rho_t} \int_0^x dy \ (1 - \rho_0 - \lambda y)p_{t-1}(y)$$
$$= \frac{1}{1 - \rho_t} [(1 - \rho_0 - \lambda x)P_{t-1}(x) + \lambda G_{t-1}(x)]$$

---

[9]The analysis in the general case is similar, however the numerical solution is computationally intense and Monte Carlo simulation is a more efficient approach.

Finally, putting it all together, we have, for $P_t(x)$

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

where $z = \frac{x - \alpha}{1 - \alpha}$ and Q_t has been defined earlier in terms of $P_{t-1}, G_{t-1}, H_{t-1}$. Note that in this notation, $q_t = (\rho_0 - \lambda)G_{t-1}(1) + 2\lambda H_{t-1}(1)$.

We then employ a dynamic program to compute $P_t(x)$, $G_t(x)$, $H_t(x)$, $Q_t(x)$ and q_t simultaneously, with initial conditions:

$$
P_0(x) = 1, \qquad G_0(x) = x, \qquad H_0(x) = \frac{1}{2}x^2.
$$

When $\alpha = 1$, the equations simplify considerably, and we have $P_1(x) = 1 - \rho_0(1 - F(x))$. For uniformly distributed X_t, $F_X(x) = x$ and we have

$$
P_1(x) = 1 - \rho_0(1 - x)
$$

$$
G_1(x) = x(1 - \rho_0(1 - \frac{1}{2}x))
$$

$$
P_t(x) = P_{t-1}(x)(1 + (\rho_0 + \lambda x)(x - 1)) + \lambda G_{t-1}(x)(1 - x)
$$

$$
G_t(x) = \int_0^x dy\ P_t(y)
$$