Collective Wisdom: Information Growth in Wikis and Blogs

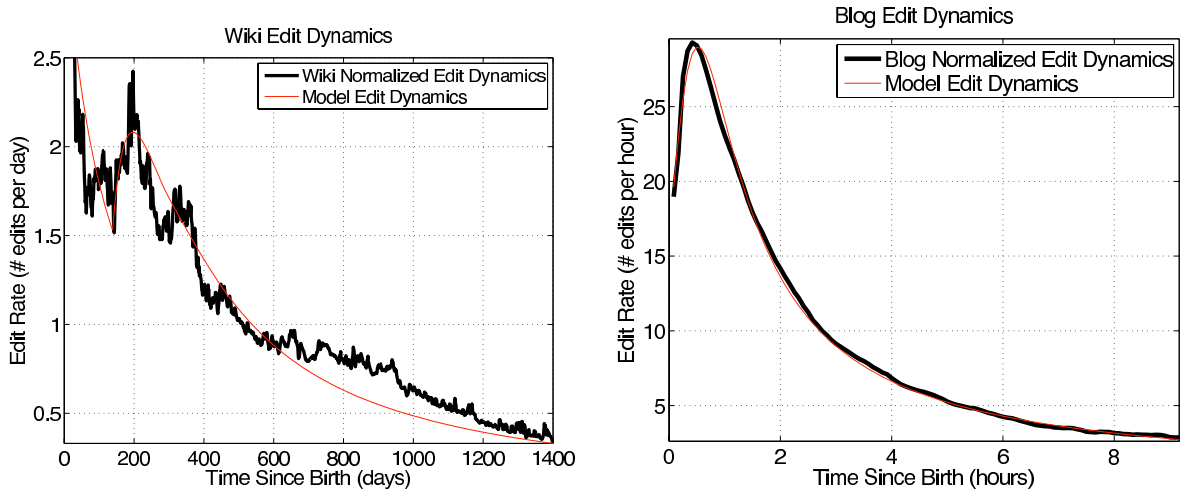
Sanmay Das and Malik Magdon-Ismail
Dept. of Computer Science, Rensselaer Polytechnic Institute

Wikis and blogs have become enormously successful media for collaborative information creation. Web users turn to blogs as sources of news and opinion and they consult Wikipedia as a reference. Wiki articles and blog posts accrue information through the asynchronous editing of users who arrive both (1) seeking information and (2) possibly able to contribute information. Most articles stabilize to reflect the collective wisdom of all the users who edited the article. The success of these media has led to new questions about the dynamic processes that create trusted sources of information – we call these *collective wisdom processes* (CWPs). We propose a model for information growth in CWPs 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). In order to validate our model we present a novel analysis of Wikipedia edit data, as well as new data from LiveJournal blogs. We show that our model is able to reproduce many features of the edit dynamics in both CWPs; in particular, it 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.

Motivation: 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. Wilkinson and Huberman (WH, 2007) may have been the first to study such dynamics. They suggest that Wikipedia follows a "rich get richer" stochastic geometric growth model in which articles accrue edits at a rate proportional to the number of edits already received. One prediction of pure rich-get-richer growth models, like the WH model, is that the total number of edits on a given wiki article or blog post should continue to increase over time. However, this does not take into account a fundamental informational limit – 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.

Data: We analyze (1) editing data for Wikipedia from its inception through May 2008, and (2) comment posting data from the Russian segment of LiveJournal from January to June of 2008. We consider all Wikipedia pages with more than 1,000 edits, and all blog posts that received more than 100 comments. We consider only the *meaningful edits* for Wikipedia pages, excluding edits attributed to vandalism or reversions of vandalism, and edits made by bots. The Wikipedia data is also normalized by the total reach of Wikipedia. The similarities in the edit rate dynamics for wikis and blogs are striking (Figure 1). 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. We can break the edit dynamics of a CWP into three regions: (1) a possible 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 in both cases). Wikis and blogs are CWPs of distinctly different natures, in terms of both content and the time-scale of editing dynamics.

Wikipedia serves as an archival resource, while blogs tend to focus on current news, providing a more conversational medium. However, the data indicates that the above stylized facts may be universal to CWPs. Any model of a CWP must then be consistent with these facts.



(a) (b)

Figure 1: Edit dynamics for (a) Wikipedia, and (b) The Russian LiveJournal blogosphere, along with the dynamics of the CWP model fit to the data in each case.

A Model: Let $t = 0, 1, \ldots$ denote the time step after the birth of a CWP. 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 $X_t \geq 0$. If $X_t > I_t$, the user has more information than is already in the CWP and the user improves the CWP. Intuitively, past visibility determines the probability of future arrivals. Visibility at a previous time step depends on the information value. If a user arrives, she may improve the quality and hence the visibility. We model ρ_t (the probability of user arrival) as a function of a base arrival probability and a visibility effect: $\rho_t = \rho_0 + \lambda V_{t-1}$. An arriving user adds some fraction α of the value she could possibly add to a CWP: if $X_t > I_{t-1}$, then the value of the CWP gets augmented to $I_t \leftarrow (1 - \alpha) I_{t-1} + \alpha X_t$. 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 blogs, the lag ℓ is close to 0 as new posts are quickly publicized through RSS feeds, while Wikipedia lags are longer, since visibility relies on search engine indexing. Note that the only observed variable is whether or not an edit occurred.

Conclusion: We develop a solution to the stochastic dynamical system implied by our model (making some standard assumptions about the distributions of random variables). This allows us to obtain the expected dynamics of the edit rate (the probability of an update at a given time step in the model). An analytical solution is possible for $\ell = 0$ (we exclude the details from this abstract). Fitting the model to the observed data from Wikipedia and LiveJournal (Figure 1) shows that this parsimonious model is sufficient to capture the exact phenomena found in the data. Not only is the fit of the model to data good, we can also use the derived fits to make nontrivial inferences about unobserved variables, like overall traffic to Wikipedia, or the amount of their knowledge that users contribute in different media.