Collective Wisdom: Information Growth in Wikis and Blogs

Sanmay Das (with Malik Magdon-Ismail)

RPI Dept. of Computer Science

“The Big Aggregators”

- Lance Fortnow, on his blog, Sep 2, 2008:

“Friday morning I wanted to know where the rumors were pointing to for McCain’s running mate selection. I could have searched various political blogs, but instead I went to Intrade and checked out the current prices on VP candidates. Since Intrade has constant trading, these prices do aggregate the various rumors and their veracity. Sarah Palin was running at about 60%. Apparently I was not the only one with this idea as Intrade had major performance problems on Friday. After seeing the price for Palin, I had a question many other Americans were asking: Who is Sarah Palin? So I went to that other great aggregator Wikipedia and read up on her.

The wisdom of crowds boiled down to a number on a trading site and a constantly updated page with much more than I need to know. The rest of the Internet is just commentary. ”

Motivation

Motivation

- Wikis and blogs are now trusted information sources

Motivation

- Wikis and blogs are now trusted information sources
- They reach stable states based on editing by users when they arrive

Motivation

- Wikis and blogs are now trusted information sources
- They reach stable states based on editing by users when they arrive
- What are the dynamics of information growth in these media?
  - Many have tried to model it as geometric growth (Wilkinson & Huberman) or as analogous to network growth (rich-get-richer)
  - But there's an informational limit!

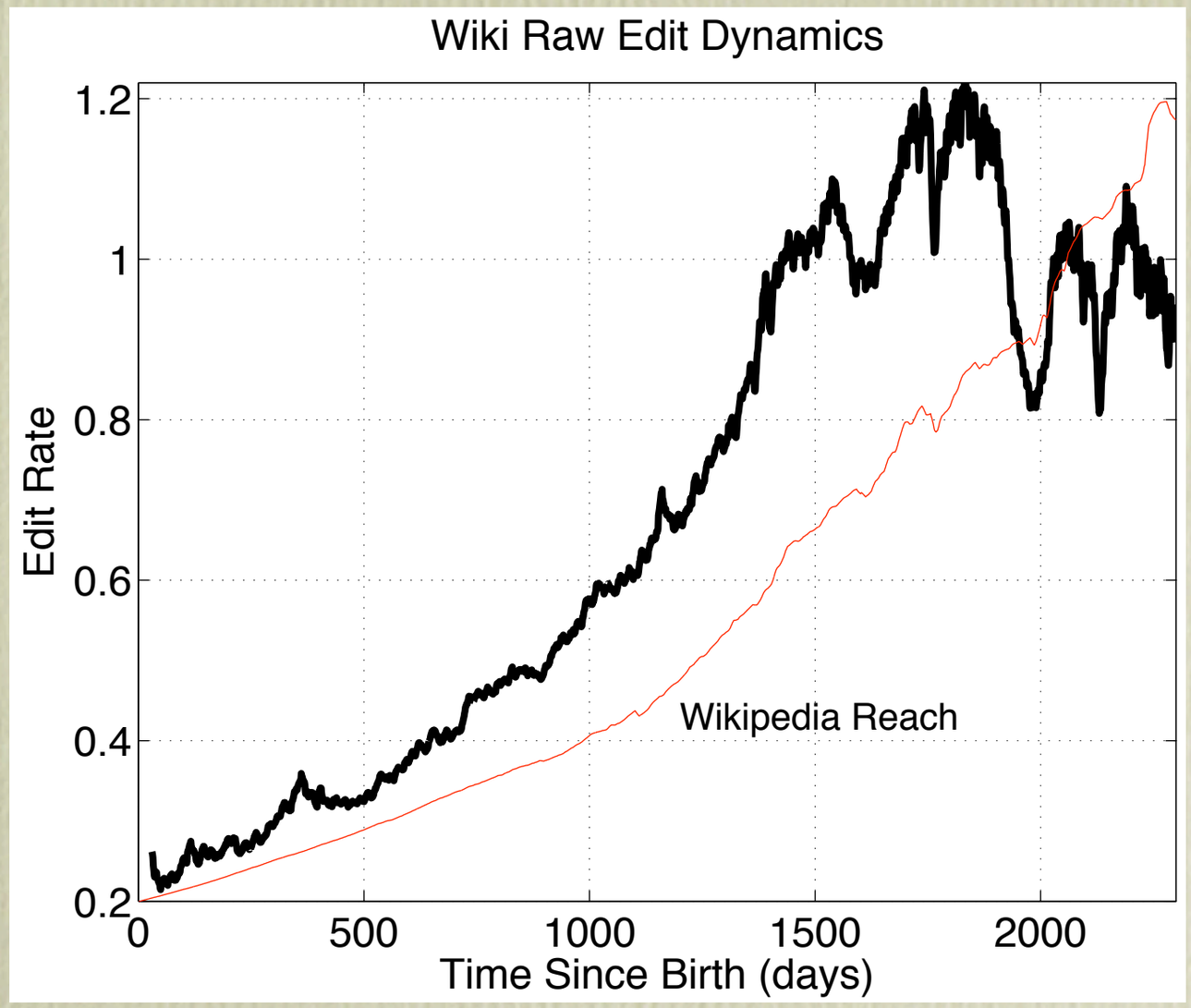
Questions

- What really happens on highly-edited Wikipedia pages?
- What about blogs? Similarities? Differences?
- Is there a good model?

Article Growth in Wikipedia

- All articles with more than 1000 edits as of May 2008 (about 15000)
- Vandalism, maintenance, and bot edits were removed
- All pages measured from inception time

Raw Edit Dynamics



The Model

- As quality improves, articles attract more visitors (**rich-get-richer**)

- As quality improves, arriving visitors are less able to contribute (**informational limit**)

Specifying the Model

$$
\begin{aligned}
V_t &= I_{t-\ell}, \quad \boxed{\text{Visibility is lagged quality}} \\
\rho_t &= \rho_0 + \lambda V_{t-1}, \quad \boxed{\text{Arrival probability}} \\
a_t &= \begin{cases} 0 & \text{w.p. } 1-\rho_t, \\ 1 & \text{w.p. } \rho_t. \end{cases}, \\
X_t &\sim F_X, \quad \boxed{\text{Knowledge of arriving user}} \\
I_t &= \max\{I_{t-1}, (1-\alpha) I_{t-1} + \alpha X_t \cdot a_t\}
\end{aligned}
$$

$\boxed{\text{State of knowledge of the page}}$

Solving the System

$$q_t = \Pr(\text{An edit occurs at time } t) = \Pr[X_t > I_{t-1}]$$

With no lag, the problem can be solved using DP

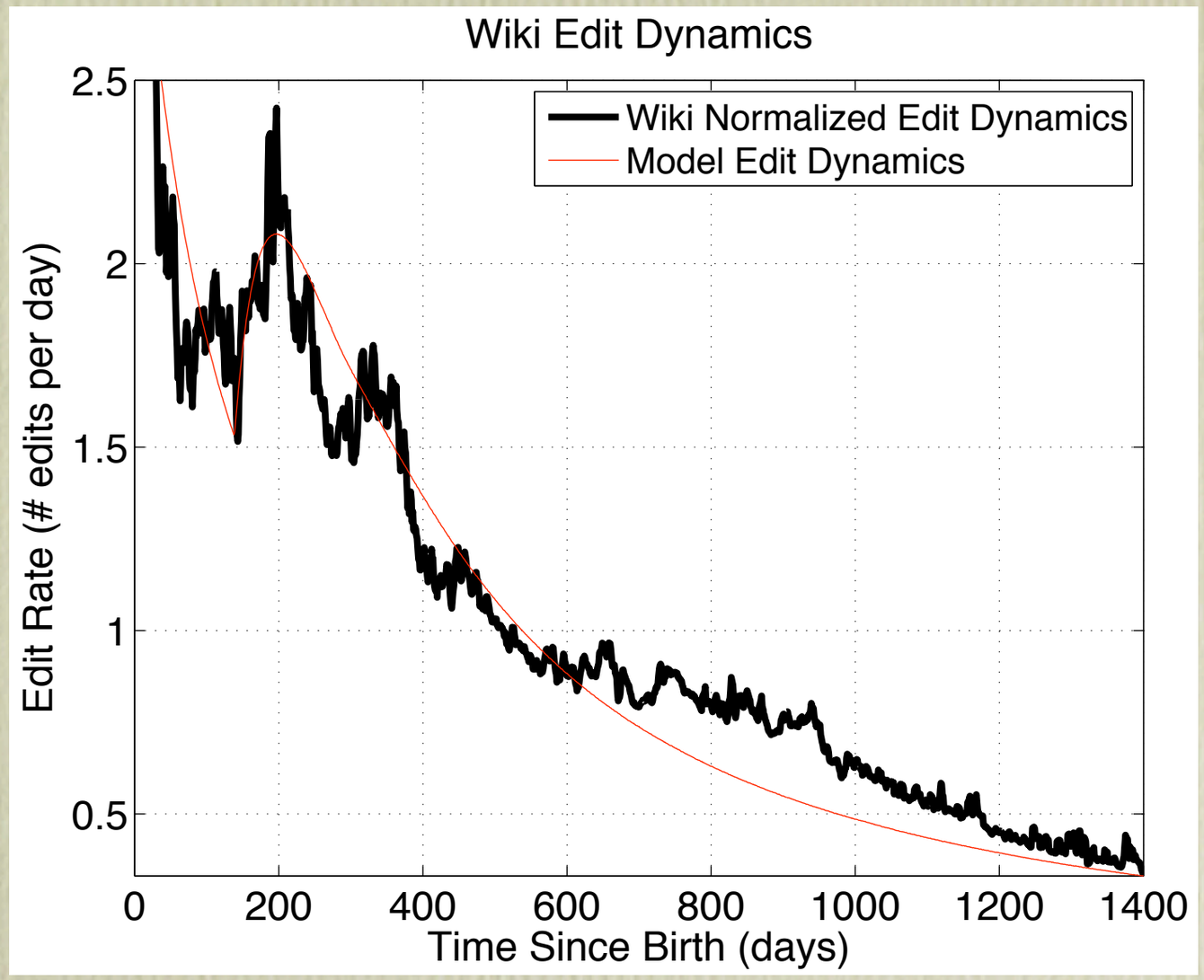
$$P_t(x) = \begin{cases} Q_t(x) + (1-\rho_0 - \lambda x)P_{t-1}(x) + \lambda G_{t-1}(x) & x \le \alpha, \\ Q_t(x) - Q_t(z) + zP_{t-1}(z)(\rho_0 + \lambda z) - \lambda z G_{t-1}(z) & \\ \quad +(1-\rho_0 - \lambda x)P_{t-1}(x) + \lambda G_{t-1}(x) & x > \alpha, \end{cases}$$

where P is the distribution function of I and $z = \dfrac{x-\alpha}{1-\alpha}$

Implications For the Edit Lifecycle

- Possible initial decline in edit rate (visibility must catch up to improvement)
- Edit rate increases to a peak with concave growth (visibility takes over)
- Edit rate decays (asymptotically as 1/t)

Wikipedia

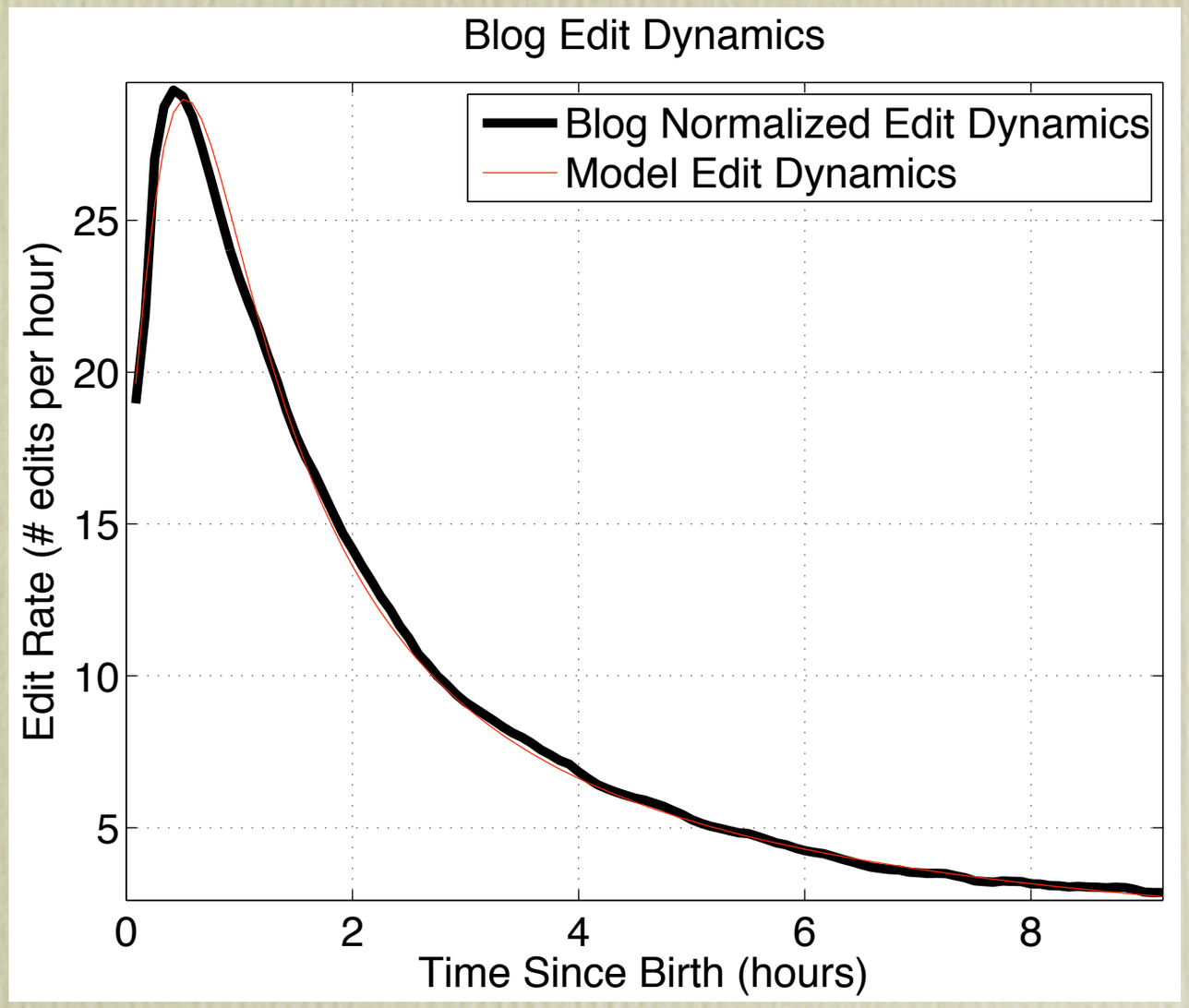

Wiki Edit Dynamics
Wiki Normalized Edit Dynamics
Model Edit Dynamics
Edit Rate (# edits per day)
2.5
2
1.5
1
0.5
0
200
400
600
800
1000
1200
1400
Time Since Birth (days)

Blogs

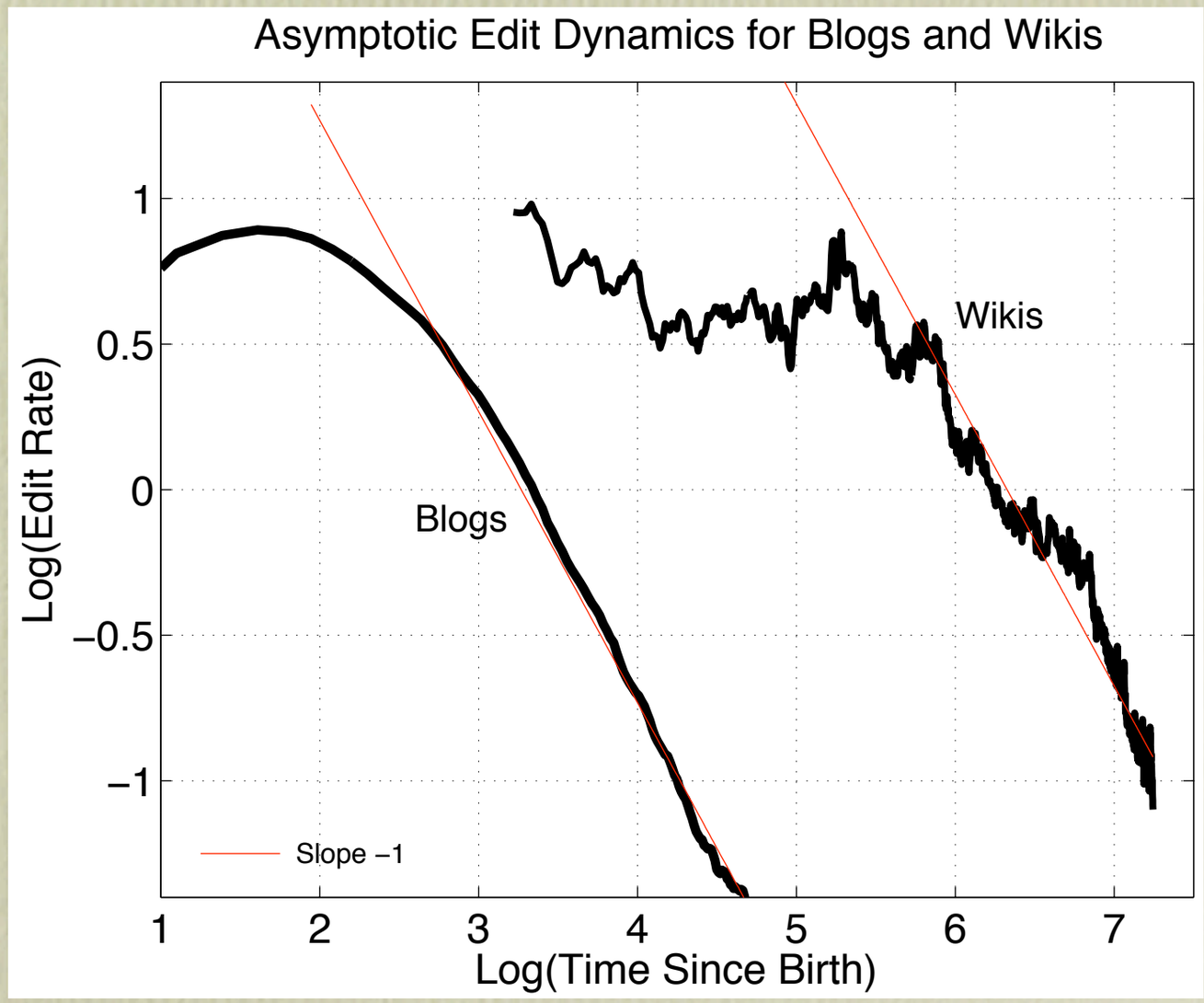
- Data collection methodology: identify posts to LiveJournal with Cyrillic characters

- Sweep through the comments to a post two weeks later, collecting all of them

Blog data


Blog Edit Dynamics
Blog Normalized Edit Dynamics
Model Edit Dynamics
Edit Rate (# edits per hour)
25
20
15
10
5
0
2
4
6
8
Time Since Birth (hours)

Decay


Asymptotic Edit Dynamics for Blogs and Wikis
Log(Edit Rate)
Log(Time Since Birth)
Wikis
Blogs
Slope −1
1
0.5
0
−0.5
−1
1
2
3
4
5
6
7

Discussion

- Model provides a good fit to data
- Can use the model to make nontrivial inferences about unobserved data:
  - People contribute less of their knowledge to Wikipedia pages than blogs
  - Overall traffic to Wikipedia would imply about 2.2 million visits to individual pages per month (real traffic ~ 1 million)

Open Questions

Open Questions

- Distribution of edits to pages when sliced at a particular cross-section? Power-law? Lognormal?

Open Questions

- Distribution of edits to pages when sliced at a particular cross-section? Power-law? Lognormal?
- Role of vandalism and edit wars in the dynamics?

Open Questions

- Distribution of edits to pages when sliced at a particular cross-section? Power-law? Lognormal?
- Role of vandalism and edit wars in the dynamics?
- Other domains: contributions to open-source software?