REMEMBERING WHAT WE LIKE: TOWARD AN AGENT-BASED MODEL OF WEB TRAFFIC

B. Gonçalves M. R. Meiss J. J. Ramasco
A. Flammini F. Menczer
MOTIVATION

- How do people navigate online?
- Can we model it effectively?
  - Applications to Ranking?
- Can we use it to predict traffic?
- Can we reconcile empirical data and theoretical models?
EMPIRICAL DATA

Dormitory (~ 1,000 users) → Campus Router → Collection System → Global Internet

Meiss et al, WSDM 2008
EMPIRICAL DATA

- \( N = 967 \) Users
- 29.8M Page requests
- 630,000 Web servers
- 110,000 Referring hosts
- 2 months of data collection Mar 5 - May 3, 2008
- MAC addresses as IDs
WEB SURFING
BOOKRANK

PRL 96: 218701, 2006

Bookmark List

Start

Jump or Browse?

1-\(p_t\)

Hit Back Button?

1-\(p_b\)

Click

Stop

Go Back

PageRank: \(p_b = 0\) No Bookmark Ranking
The graph shows the probability distribution of jumps (P(Jump)) on a log-log scale. The click stream data forms a straight line with a slope of 1.8, indicating a power-law distribution. The line graph is overlaid with a trend line that best fits the data points, highlighting the linear relationship between the frequency of jumps and their magnitude.
SITE TRAFFIC BOOKMARKS

\[ P(Jump) \]

PageRank

slope = 1.8
BOOKMARK TRAFFIC

\[ P(Jump) \]

- click stream
- BookRank
- PageRank
- slope = 1.8
SITE TRAFFIC

P(T) vs T plot showing the relationship between page traffic and time. The graph includes lines labeled as click stream, BookRank, PageRank, and a green line indicating a slope of 1.8.
LINK TRAFFIC

\[ P(\omega) \]

- click stream
- slope = 1.9
- BookRank
- PageRank
SHANNON ENTROPY

Definition

\[ S = -\sum_i \rho_i \log \rho_i \]

- \( S = 0 \) All visits are to same site
- \( S = \log n \) One visit to each site
- Measures information needed to describe a user browsing pattern
ENTROPY DISTRIBUTION

\[ P(S) \]

- click stream
- PageRank
ENHEDY DISTRIBUTION

\[ P(S) \]

click stream
PageRank
\[ p_b = 0.01 \]
ENTROPY DISTRIBUTION

- Click stream
- PageRank

$p_b = 0.01$

$p_b = 0.2$
ENTROPY DISTRIBUTION

- Click stream
- PageRank
- $p_b = 0.01$
- $p_b = 0.2$
- $p_b = 0.4$
TIME BETWEEN VISITS

\[ P(\tau) \]

\( \tau \) (clicks)
DISCUSSION

- PR does not predict real traffic

- Real users are less diverse than random walkers
  - Focused interests and recurring habits

- BR adds well known user behaviors:
  - bookmarks and backtracking

- BR reconciles individual behaviour and aggregate patterns

- BR improves PRs predictions on several empirical measures
FUTURE WORK

- Multiple tabs
- User diversity, topics of interest
- Site dependent jump probability
- Different parameter values
BOOKRANK

- Add node to Bookmark list
- Jump to bookmark with prob $p_t$. $P(R) \sim R^\beta$
  - Bookmarks ranked by traffic
- With prob $1-p_t$ navigate locally
  - Prob $p_b$ press back button
  - Prob $1-p_b$ follow random link

PRL 96: 218701, 2006
REFERRALS PER HOST

![Graph showing the distribution of unique referrers per unique host. The x-axis represents the log scale of unique referrers/unique hosts ranging from $10^{-1.2}$ to $10^{0.0}$, and the y-axis represents the probability density function $P(\text{UR} / \text{UH})$ ranging from 0 to 3.5. The graph includes two curves: one for Portal Users (dashed line) and another for Surfers (dotted line).]
USERS PER REFERRAL

$P(\text{Users}) = 10^{-7} \cdot 10^{-6} \cdot 10^{-5} \cdot 10^{-4} \cdot 10^{-3} \cdot 10^{-2} \cdot 10^{-1} \cdot 10^0 \cdot 10^1 \cdot 10^2 \cdot 10^3$

exponent = 1.9
USERS PER HOST

Users

exponent = 1.9
INTERCLICK TIME

![Graphs showing the relationship between P(Interval) and Interval (sec) for two different values of 1.45 and 1.6.]