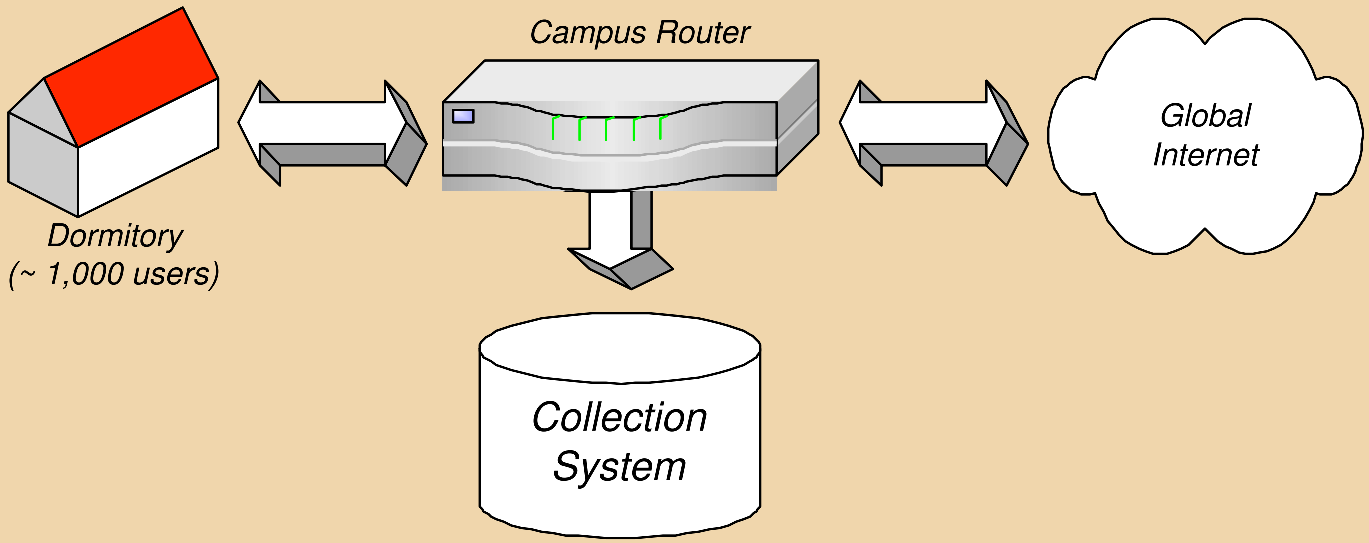
REMEMBERING WHAT WE LIKE: TOWARD AN AGENT-BASED MODEL OF WEB TRAFFIC

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

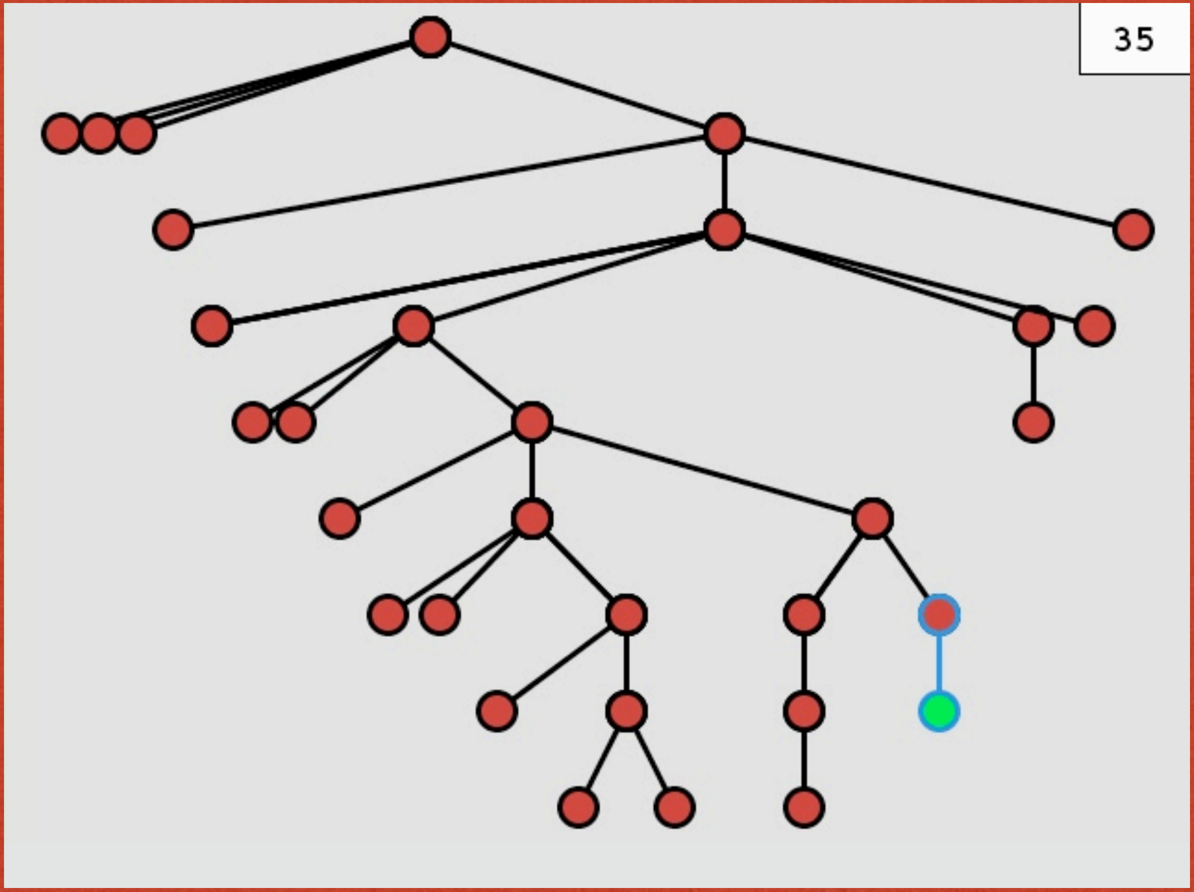


Meiss *et al*, WSDM 2008

EMPIRICAL DATA

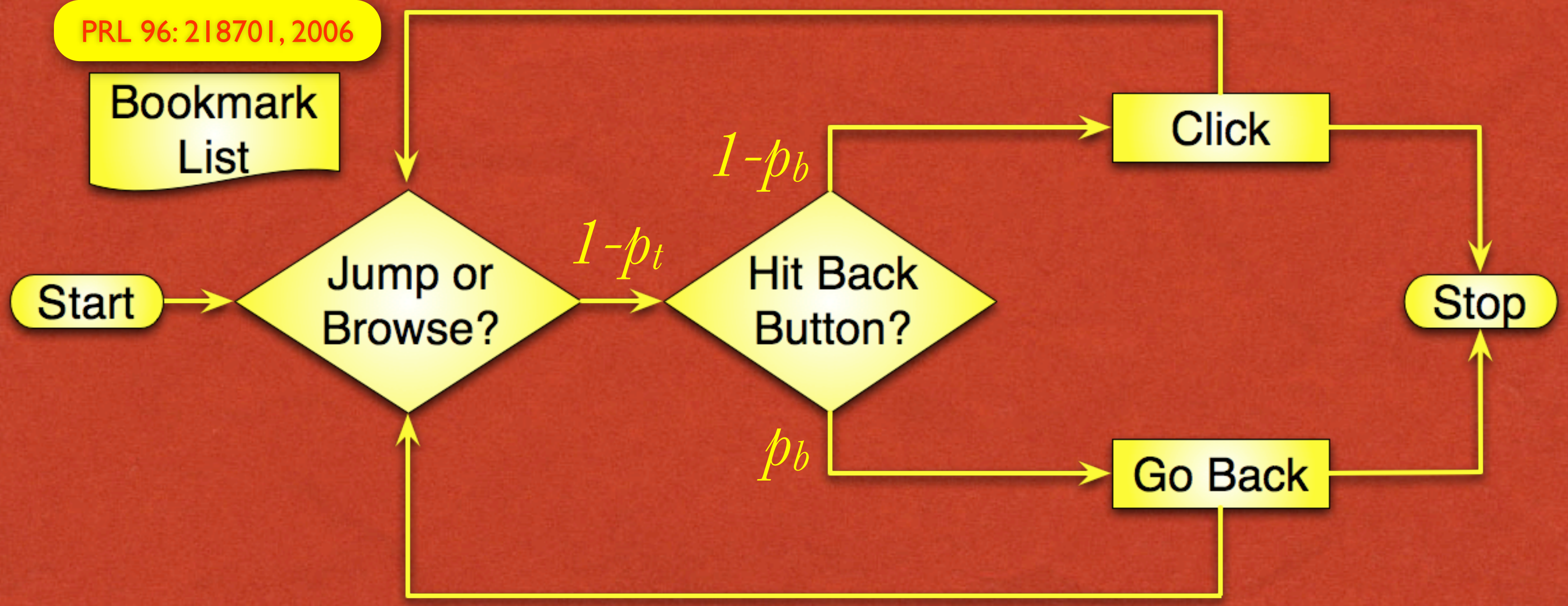
- $N = 967$ Users
- 29.8 M Page requests
- 630,000 Web servers
- 110,000 Referring hosts
- 2 months of data collection Mar 5 - May 3, 2008
- MAC addresses as IDs

WEB SURFING



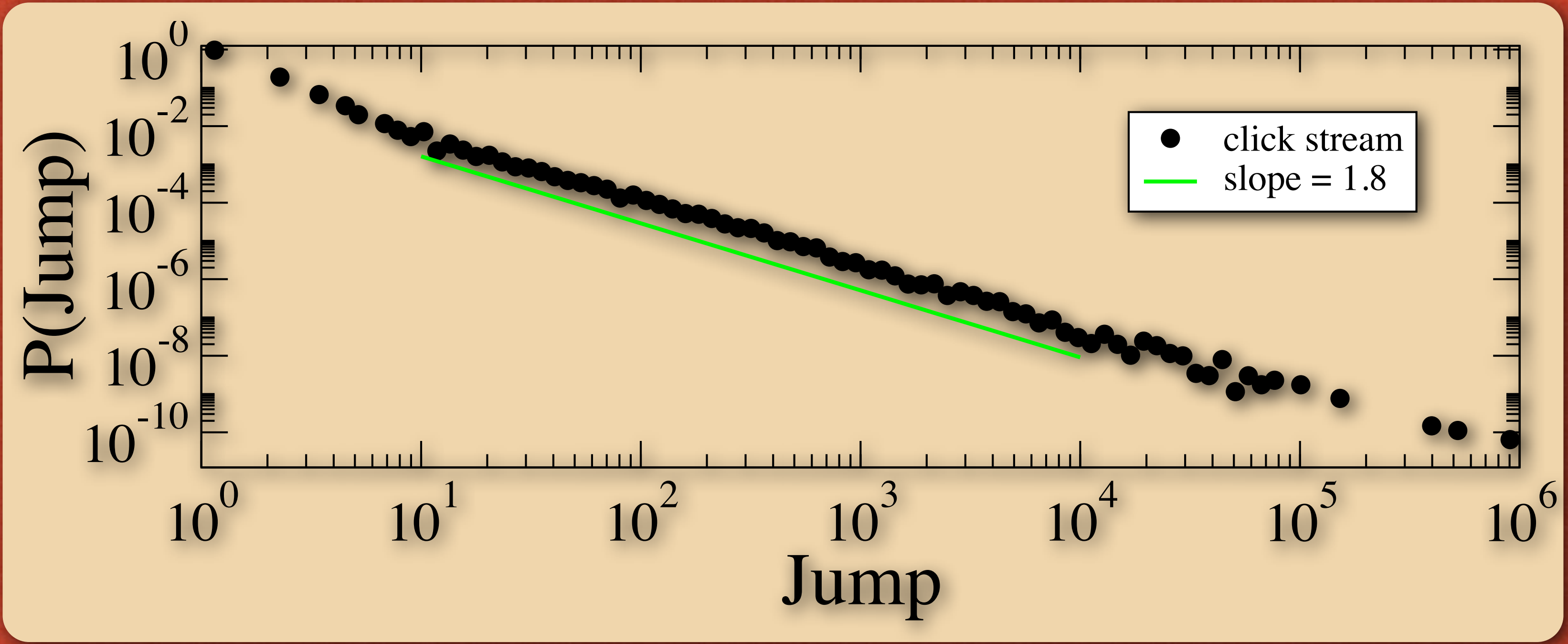
BOOKRANK

PRL 96: 218701, 2006

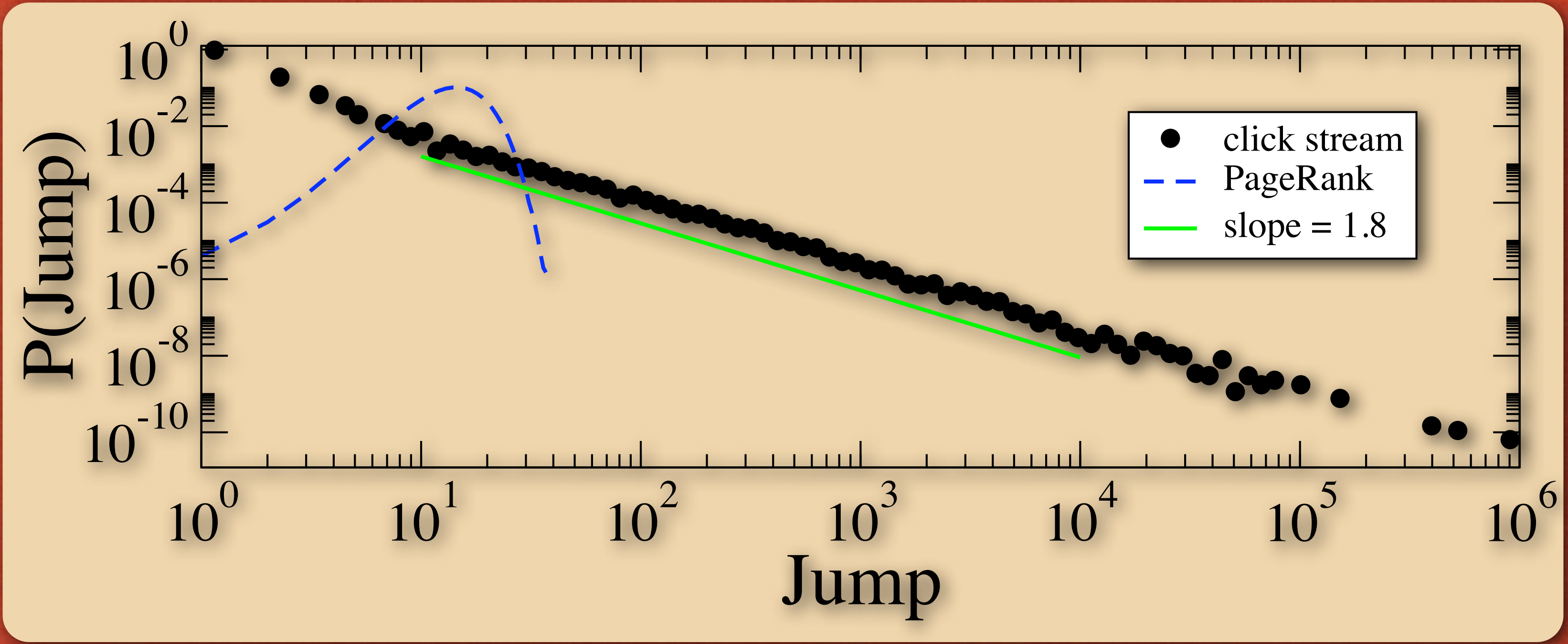


PageRank: $p_b=0$ No Bookmark Ranking

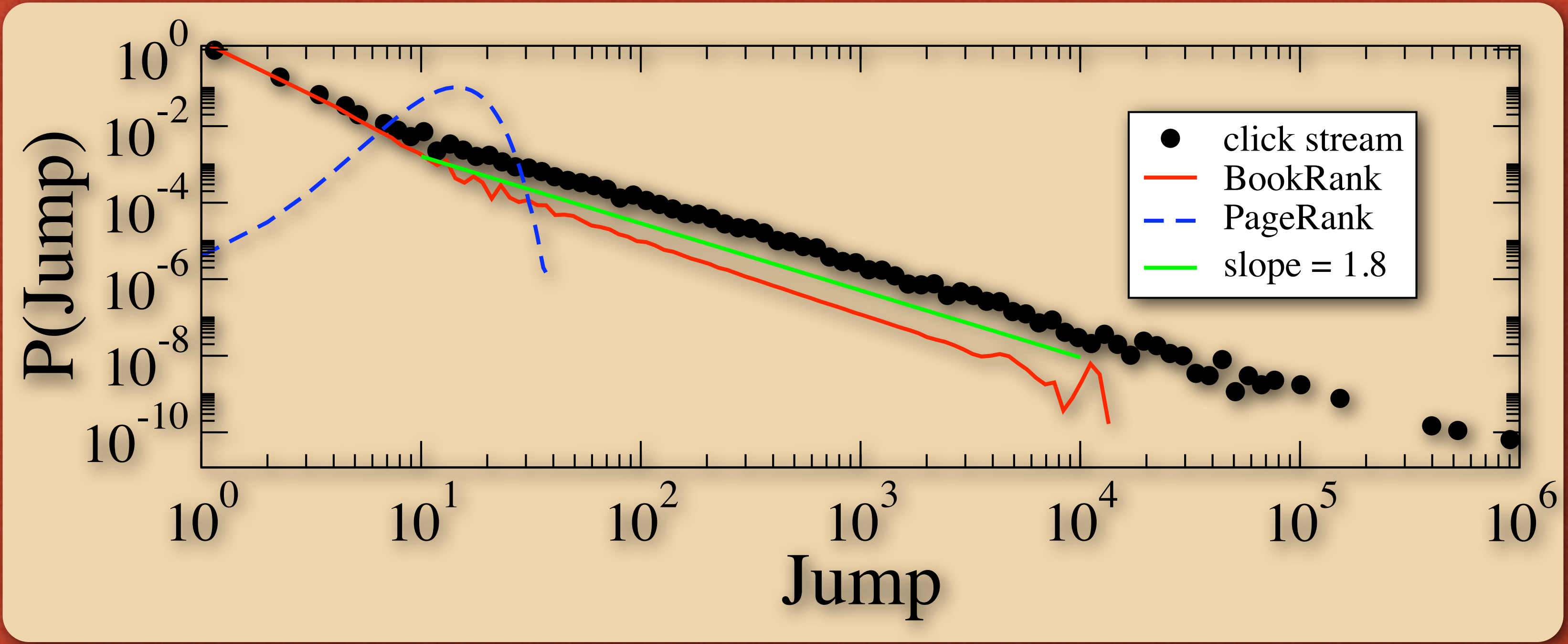
SITE TRAFFIC BOOKMARKS



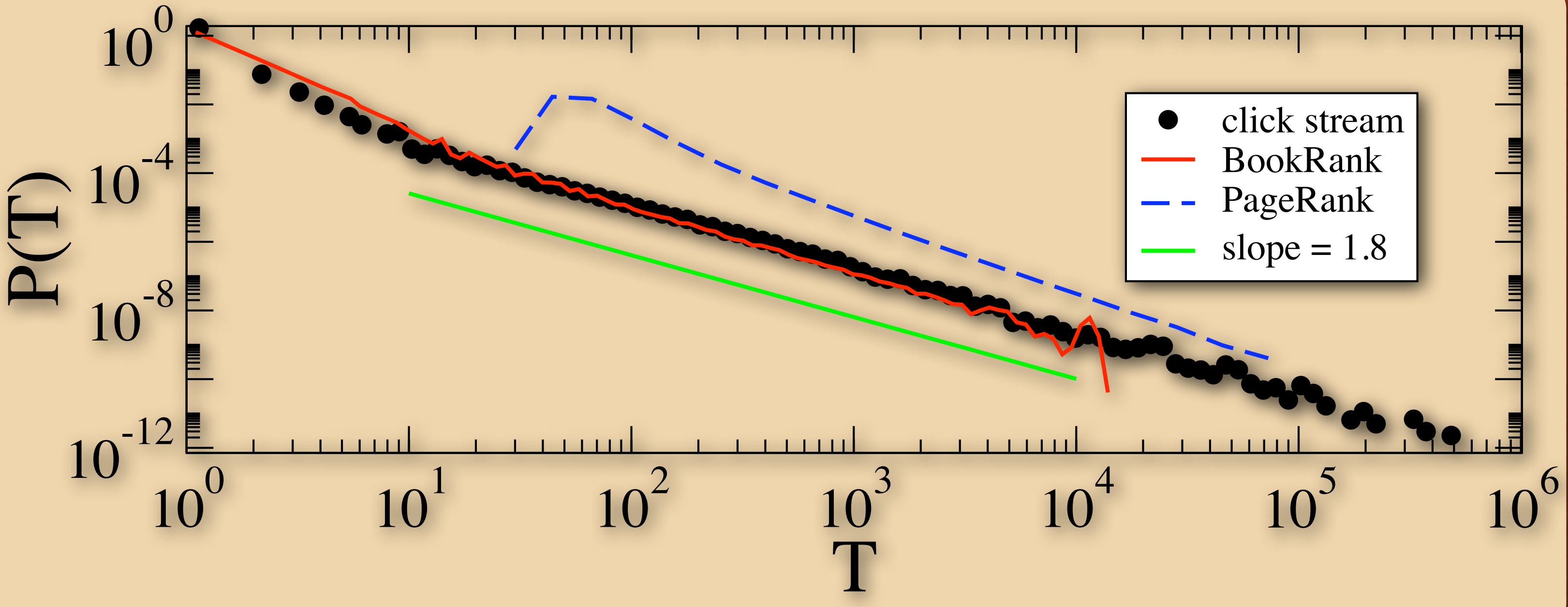
SITE TRAFFIC BOOKMARKS



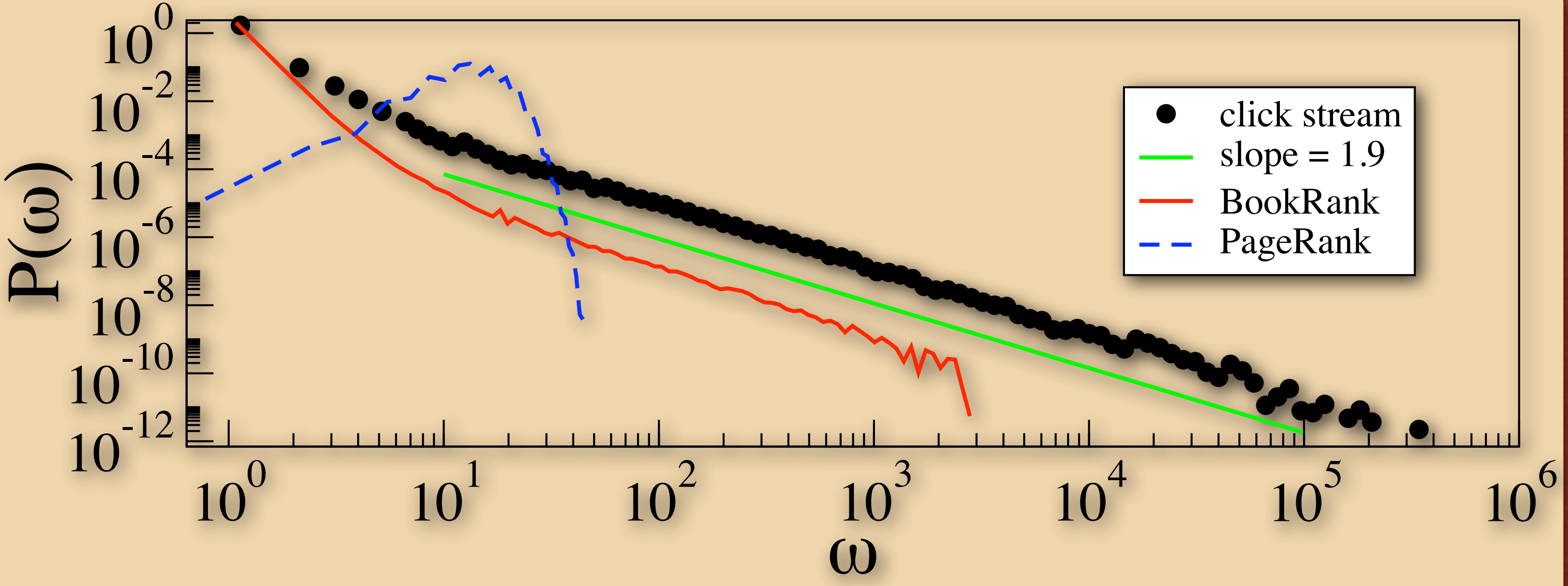
BOOKMARK TRAFFIC



SITE TRAFFIC



LINK TRAFFIC



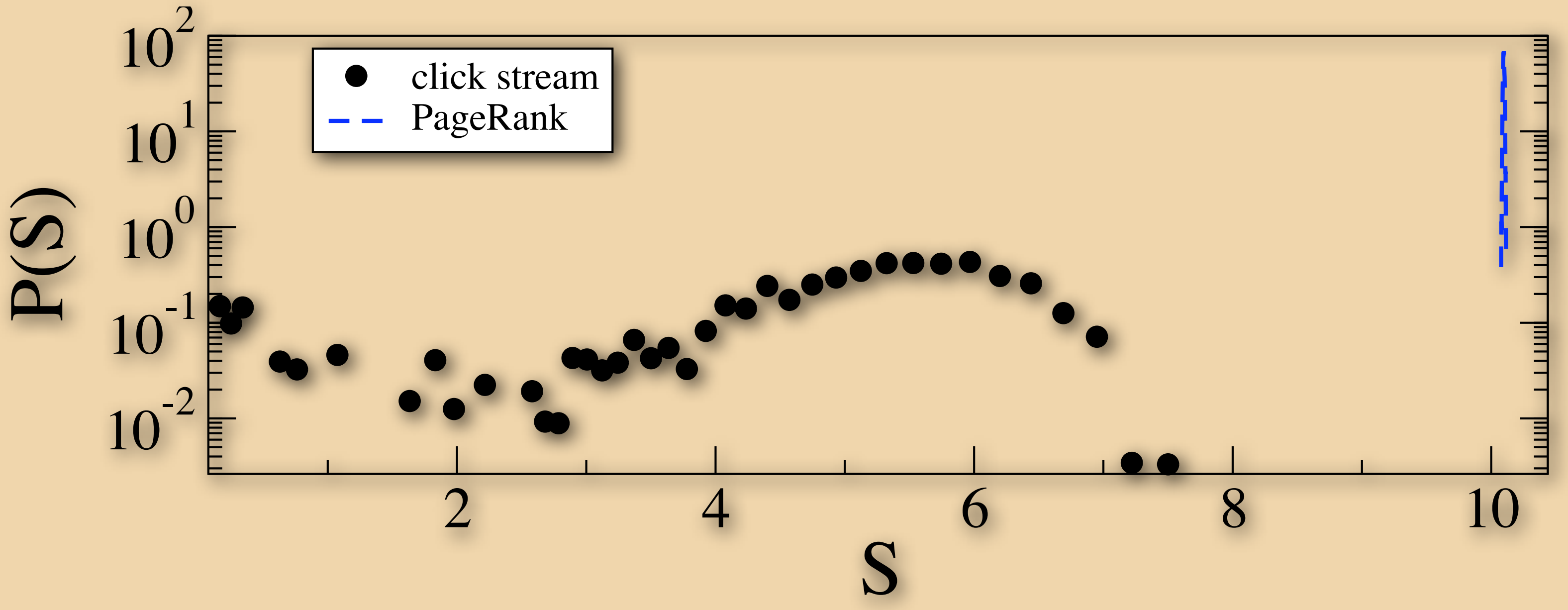
SHANNON ENTROPY

- Definition

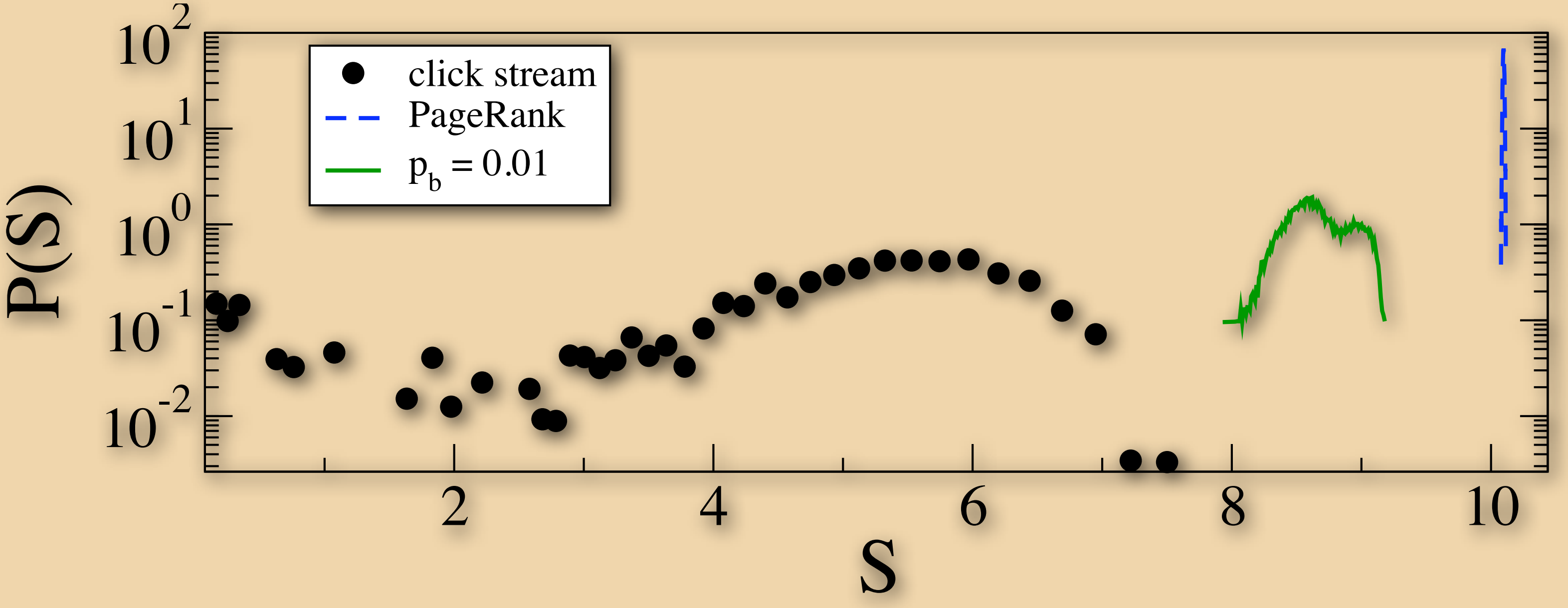
$$S = -\sum_i \rho_i \log \rho_i$$

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

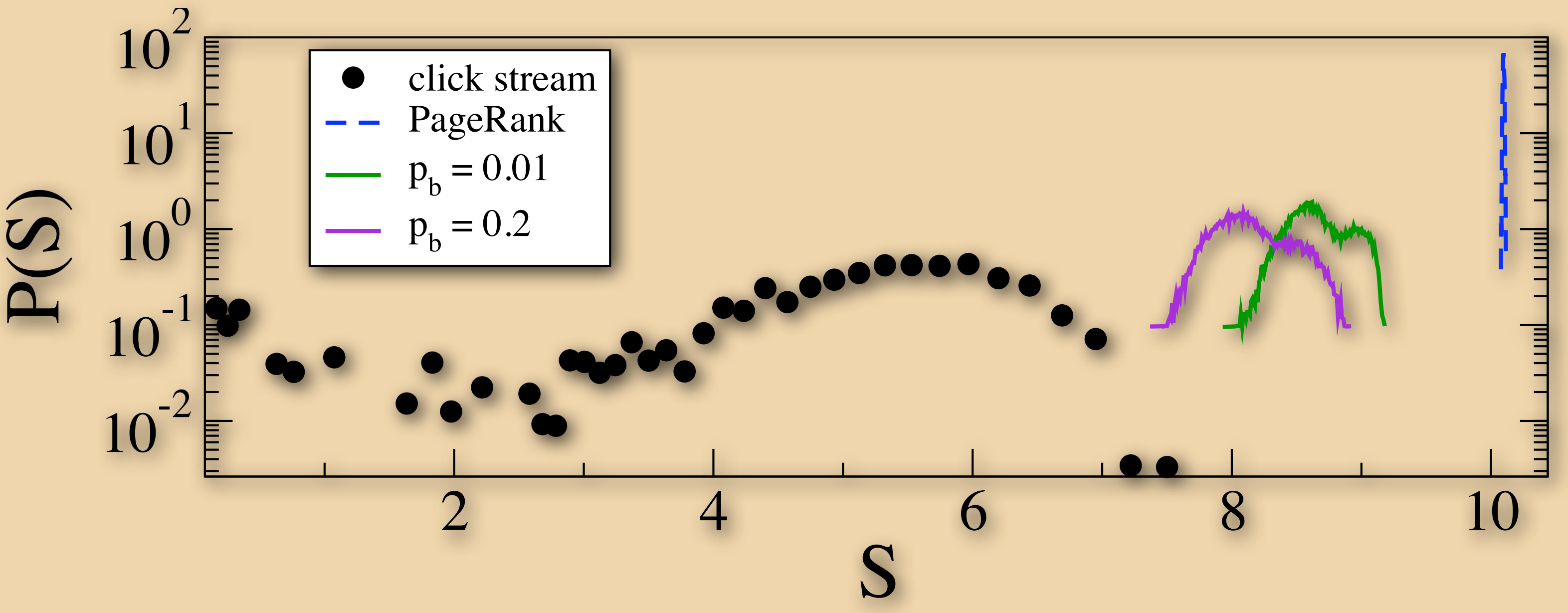
ENTROPY DISTRIBUTION



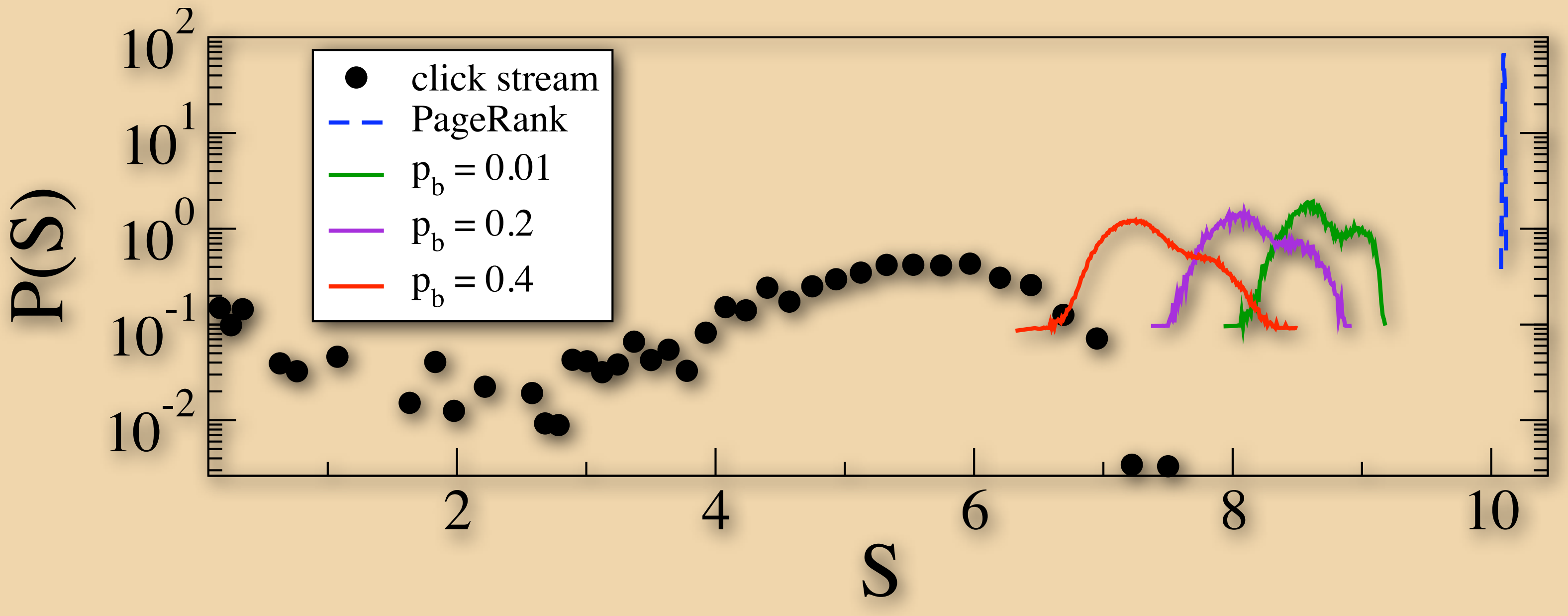
ENTROPY DISTRIBUTION



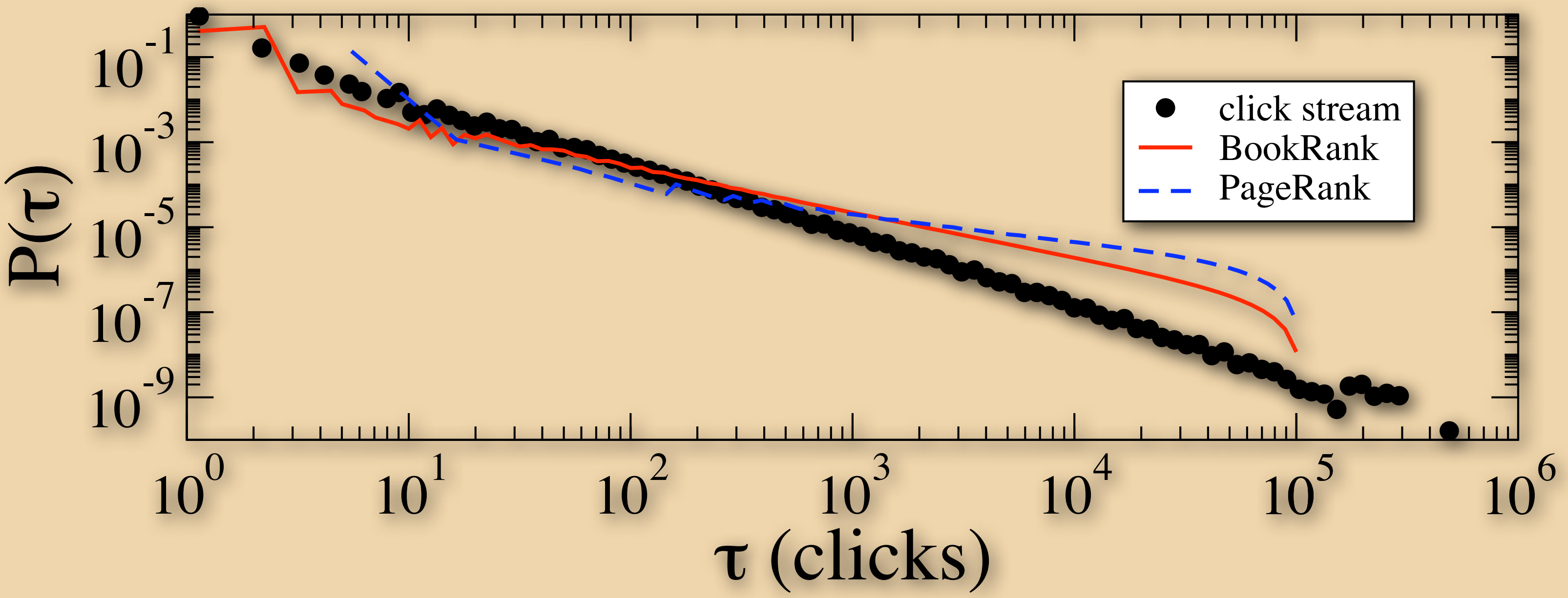
ENTROPY DISTRIBUTION



ENTROPY DISTRIBUTION



TIME BETWEEN VISITS



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

Kirkw
Indiana

Answers
NEXT EXIT ↗

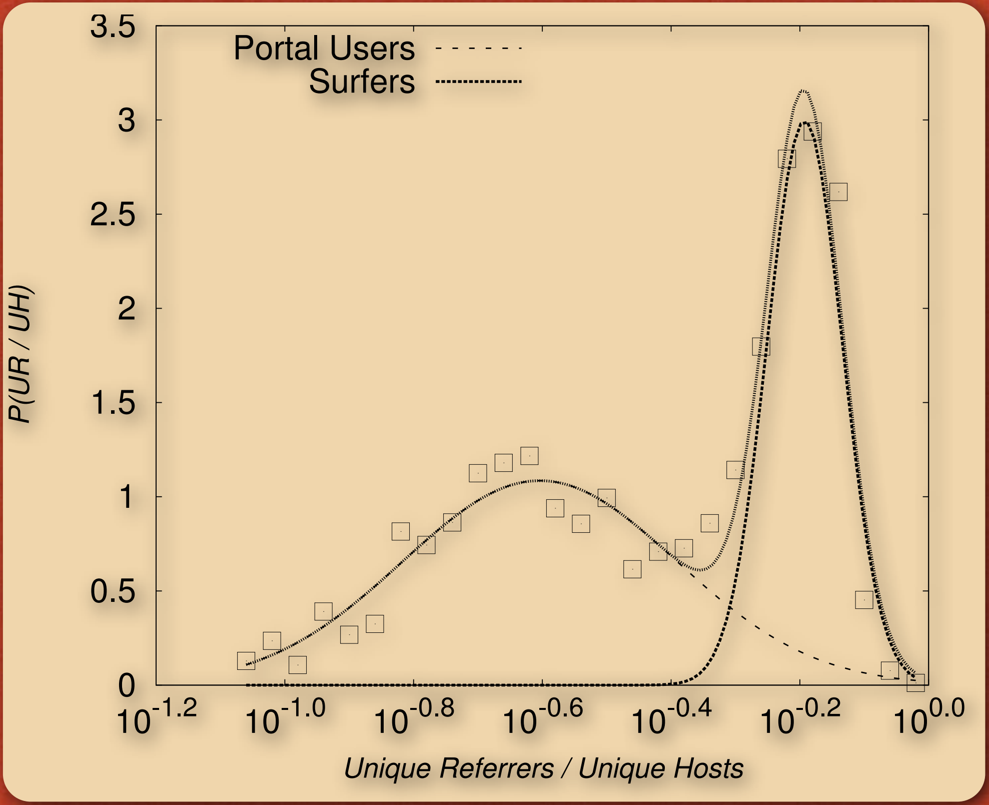
BOOKRANK

- Add node to Bookmark list
- Jump to bookmark with prob p_t. $P(R) \sim R^{\beta}$
  - Bookmarks ranked by traffic

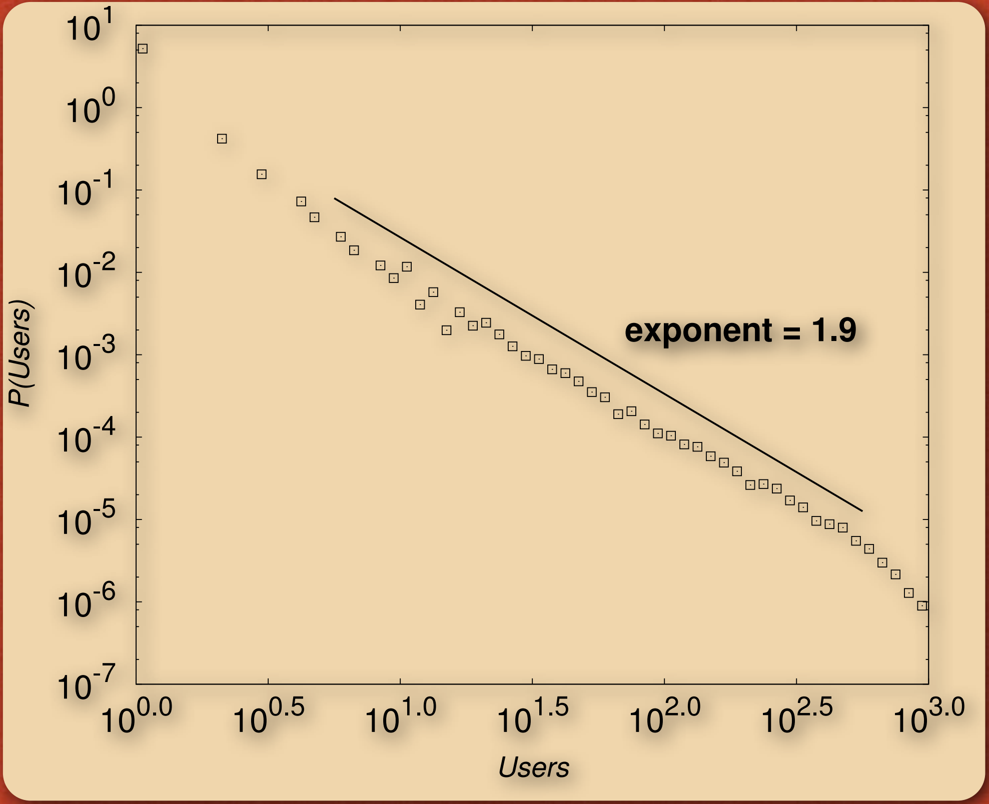
PRL 96: 218701, 2006

- With prob $1\text{-}p_t$ navigate locally
  - Prob p_b press back button
  - Prob $1\text{-}p_b$ follow random link

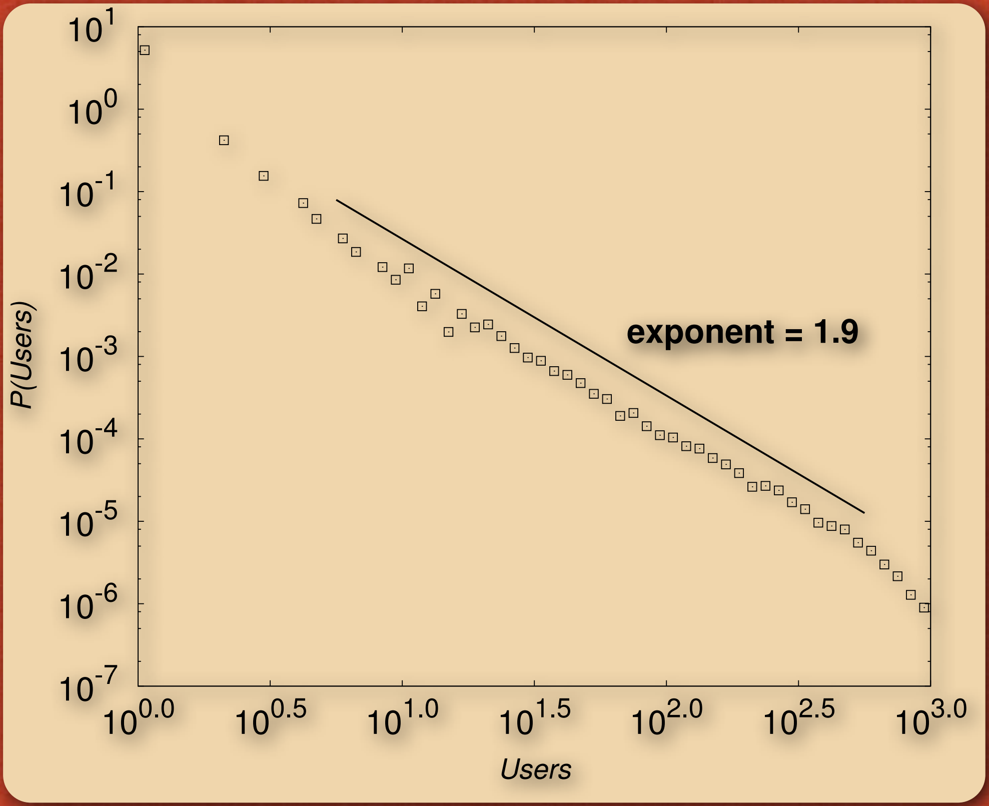
REFERRALS PER HOST



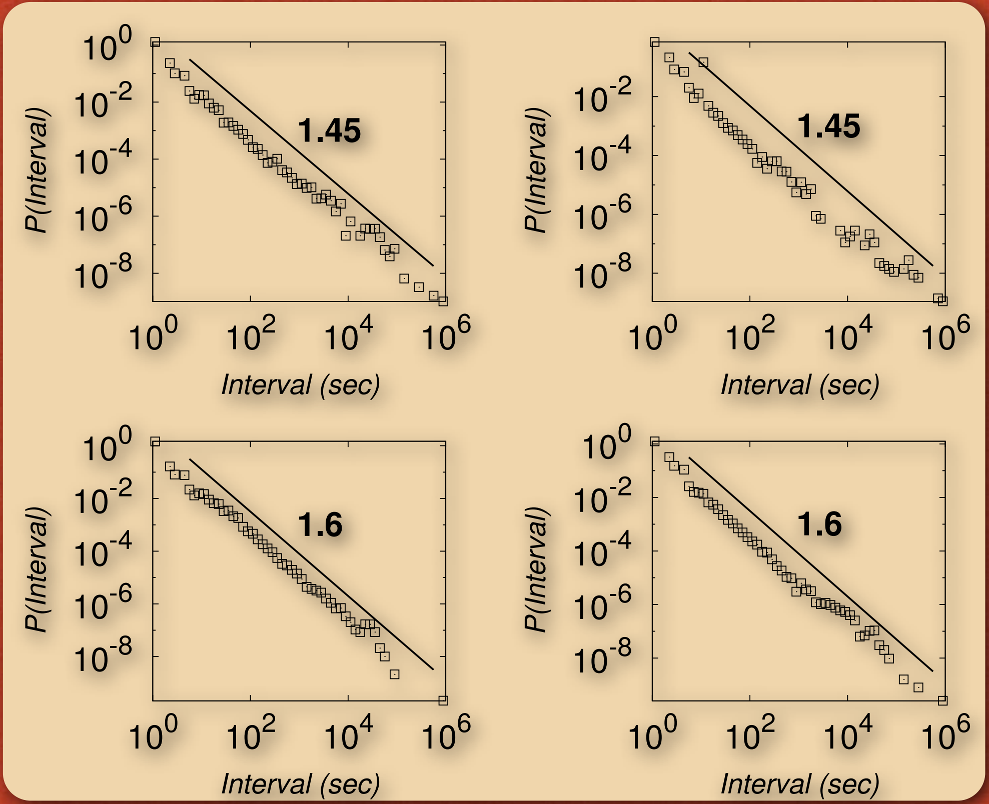
USERS PER REFERRAL



USERS PER HOST



INTERCLICK TIME


10^0
10^{-2}
10^{-4}
10^{-6}
10^{-8}
P(Interval)
10^0
10^2
10^4
10^6
Interval (sec)
10^{-2}
10^{-4}
10^{-6}
10^{-8}
P(Interval)
10^0
10^2
10^4
10^6
Interval (sec)
10^0
10^{-2}
10^{-4}
10^{-6}
10^{-8}
P(Interval)
10^0
10^2
10^4
10^6
Interval (sec)
10^0
10^{-2}
10^{-4}
10^{-6}
10^{-8}
P(Interval)
10^0
10^2
10^4
10^6
Interval (sec)