Measurement and Analysis of an Online Content Voting Network: A Case Study of Digg

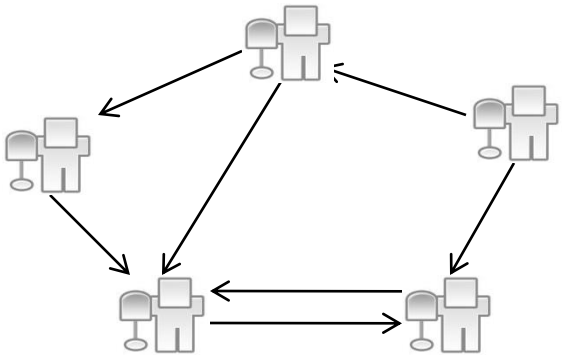
YINGWU ZHU
SEATTLE UNIVERSITY
EMAIL: ZHUY@SEATTLEU.EDU

What are online content voting networks?

- Examples:
  - Digg (stories), YouTube (videos), Flickr (photos)
- Built on an underlying social network
- Users submit and rate content
  - Popularity and availability of (UGC) content are driven by user participation
  - UGC: unprecedented scale, high dynamics, divergent quality

Background: Digg (1)

- www.digg.com
- A popular news aggregator site
- Built on an underlying social network
  - Friend links (outgoing links)
  - Fan links (incoming links)



Background: Digg (2)

- Two sections to place content
  - Upcoming stories: newly submitted stories
  - Popular stories (front page): promoted stories
    - High volume of visits (several million visits per day)
    - Can bring profits (advertisement)
- Content promotion: upcoming → front page
  - User diggs/votes
- Content filtering by two filters
  - *Friends* interface: tracks one' friends' activities
  - Front page: displays popular stories

This work

- Presents *large-scale measurement study and analysis* of the online content rating network, Digg
  - Over 52 months worth of digg trace data
- Our goals
  - Understand structural properties of Digg social network
  - Examine user digg activities
  - Explore impact of the social network on user digg activities, content promotion and content filtering

Why study online content voting networks?

- UGC is reshaping the Internet landscape
  - Web sites provides facilities to publish UGC
  - Users are publishers, consumers and referees
- User participation makes high-quality content thrive
- Technical challenges
  - Content promotion
    - Promote high-quality content
    - Profits from high volume of visits
    - Resilient to system gaming
  - Content filtering
    - Presents high-quality, interesting content to users
    - Helps users in content discovery

Rest of the talk

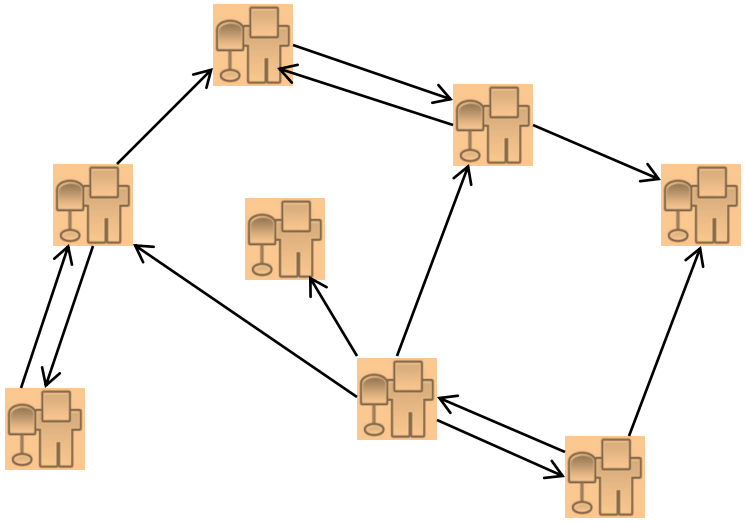
- Analyzing structural properties of Digg social network
- Measuring user digg activities
- Understanding impact of social network on user diggs, content promotion and content filtering

Crawl of social graph

- Use Digg APIs, subject to rate-limiting
- Pick known seed user "kevinrose"
  - Crawled all of his friends and fans
  - Add new users to the list
  - BFS traversal
- Continued until the list is exhausted
  - 3/10/2009 – 3/16/2009
  - WCC of the social graph

Crawl of social graph

- Use Digg APIs, subject to rate-limiting
- Pick known seed user "kevinrose"
  - Crawled all of his friends and fans
  - Add new users to the list
  - BFS traversal
- Continued until the list is exhausted
  - 3/10/2009 – 3/16/2009
  - WCC of the social graph



Crawl of user diggs

- Use Digg APIs, subject to rate-limiting
- For each crawled user, fetch his/her diggs
- Two digg traces
  - PT: spanning 2004/12/01 – 2009/03/16
  - ST: spanning 2009/03/17 – 2009/04/16
    - Study impact of the social graph on user diggs due to its recency
    - The underlying social graph did not change much over the duration of ST

High-level data characteristics

Data	Value
# of users in WCC	580, 228
# of friend links in WCC	6, 757, 789
Avg # of friend links per user	11.65
# of diggs in PT	154,129,256
Avg # of diggs per user in PT	265
Frac. of diggs submitted by WCC	90.75%
# of submitted stories in ST	257,536
# of popular/promoted stories in ST	4,571
Frac. of users in WCC dugg in ST	0.22

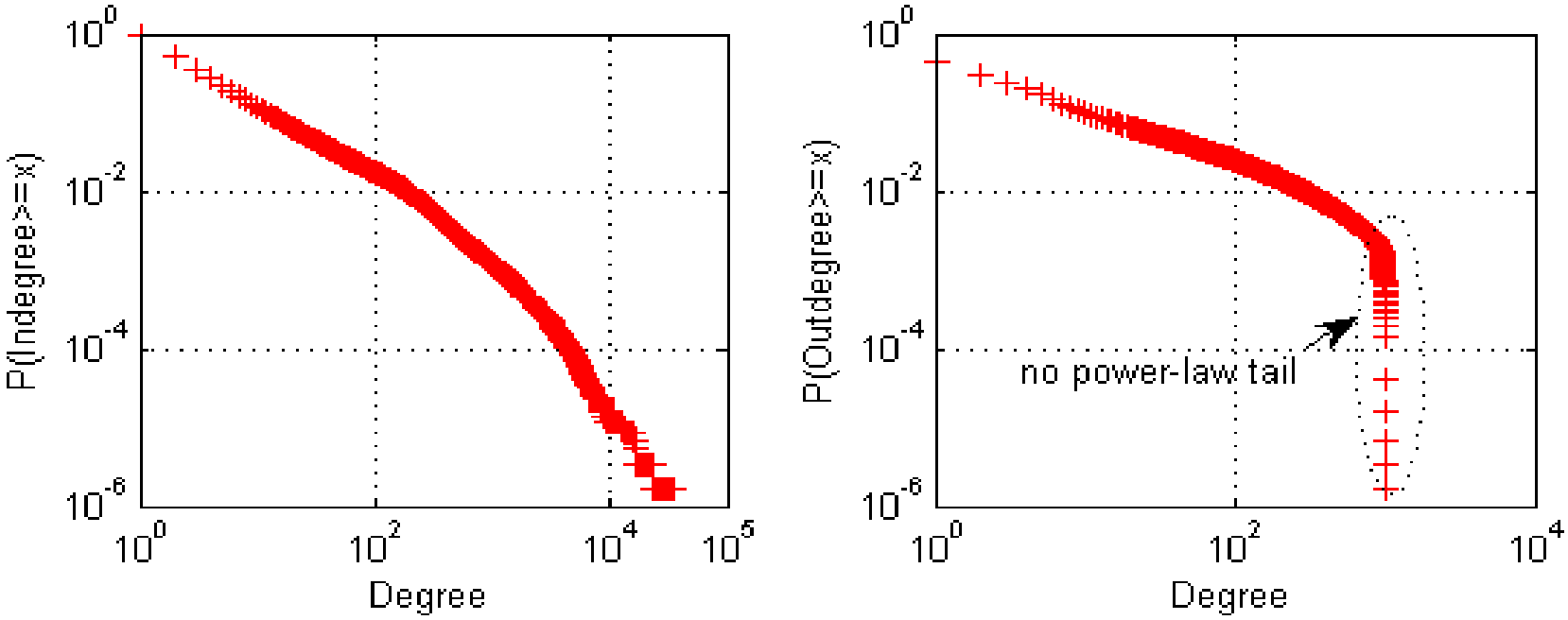
Social graph questions

- Want to examine structural properties
- How does Digg social network differ from other online social networks (OSN)?
  - Such as YouTube, Flickr, LiveJournal in prior studies [IMC07]

Link symmetry

- Digg has low link symmetry: 39.4%
  - Other OSNs show high link symmetry
  - YouTube, Flickr, LiveJournal, Orkut, Yahoo!360: 62-100%
- Speculate that Digg users are centered on story submission & rating instead of reciprocating users with social links
- Exploit low link symmetry to identify reputed digg users
  - The Web graph has low link symmetry, which is exploited by PageRank to identify trusted Web pages

CCDF of node degree distribution



1. Other OSNs' node degree shows a power-law distribution, e.g., [IMC07]
2. Digg's node out-degree distribution does not have a power-law tail
   - Low link symmetry
   - Digg users rely on story submission & voting to boost their profiles instead of aggressively creating friend links

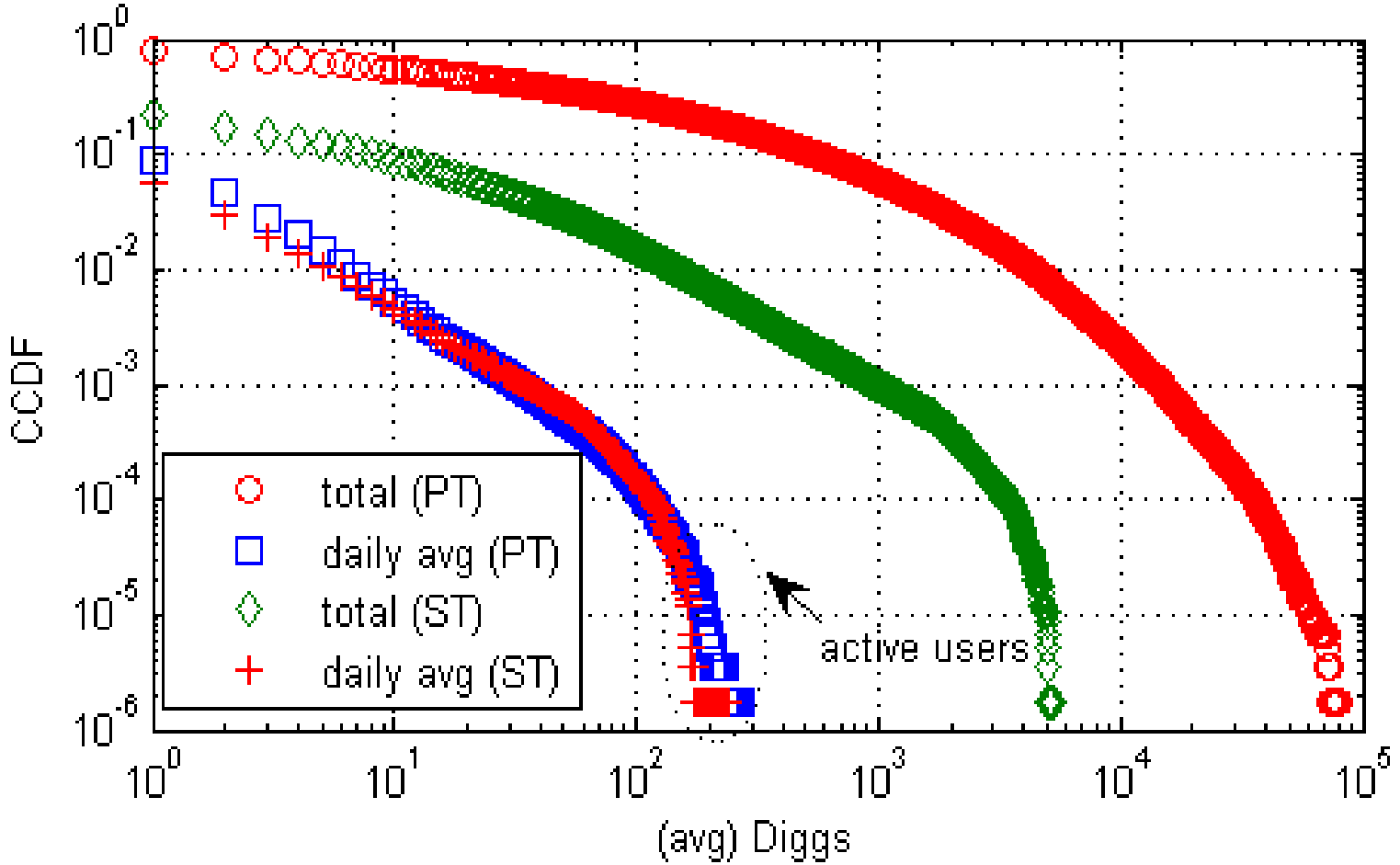
Other structural properties

- Digg exhibits weaker correlation of indegree and outdegree
  - 58% overlap for top 1% of nodes ordered by in- and outdegree, due to *low link symmetry*
  - YouTube, Flickr, LiveJournal: stronger, *nodes with high outdegree tend to have high indegree (overlap >= 65%)*
- Digg nodes tend to connect to nodes with very different degree of their own
  - Flickr, Orkut, LiveJournal: a tendency of higher degree nodes to connect to other high degree nodes
- Clustering (coefficient = 0.218)
  - Measures connection density of the neighborhood of a node
  - *coeff = # of links between friends / # of links that could exist*
  - YouTube, Flickr, Orkut, LiveJournal: 0.136 – 0.330

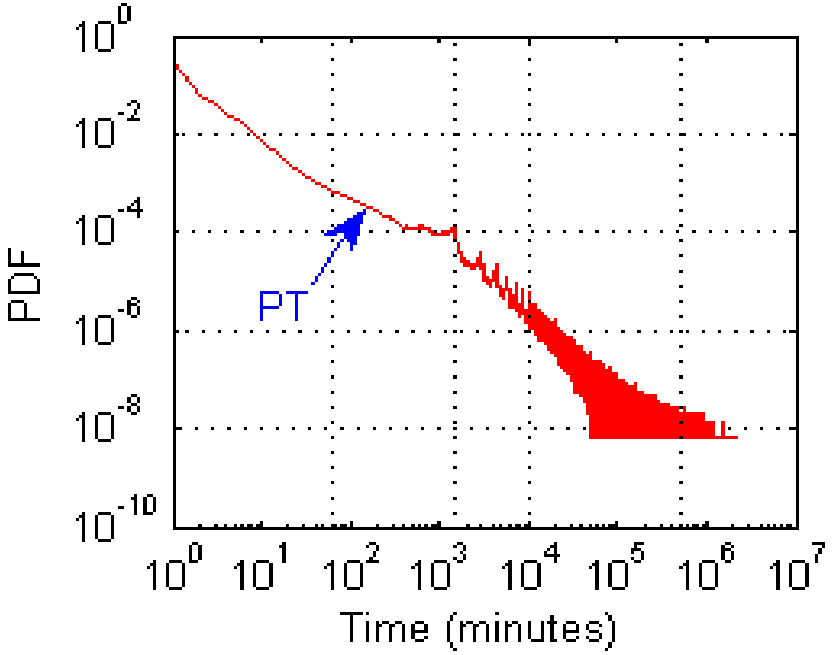
Outline

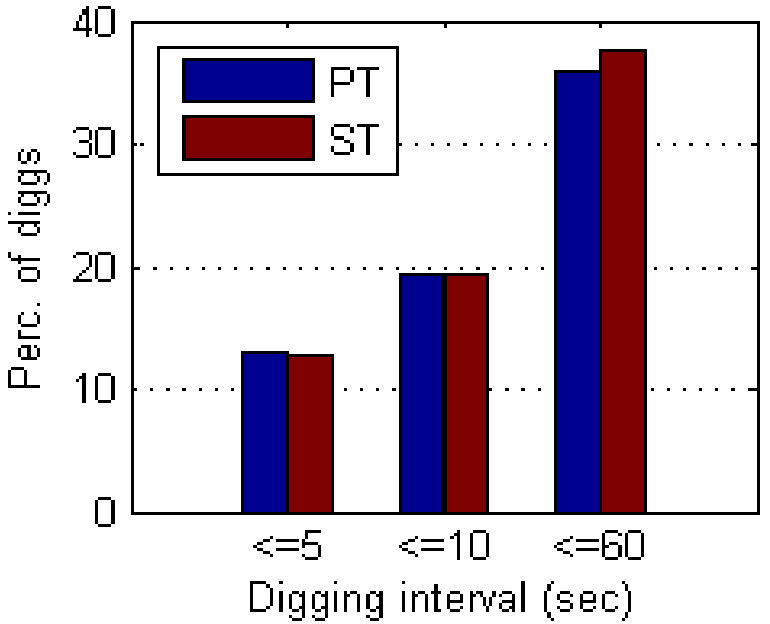
- ~~Analyzing structural properties of Digg social network~~
- Measuring user digg activities
- Understand impact of social network on user diggs, content promotion and content filtering

CCDF of user diggs



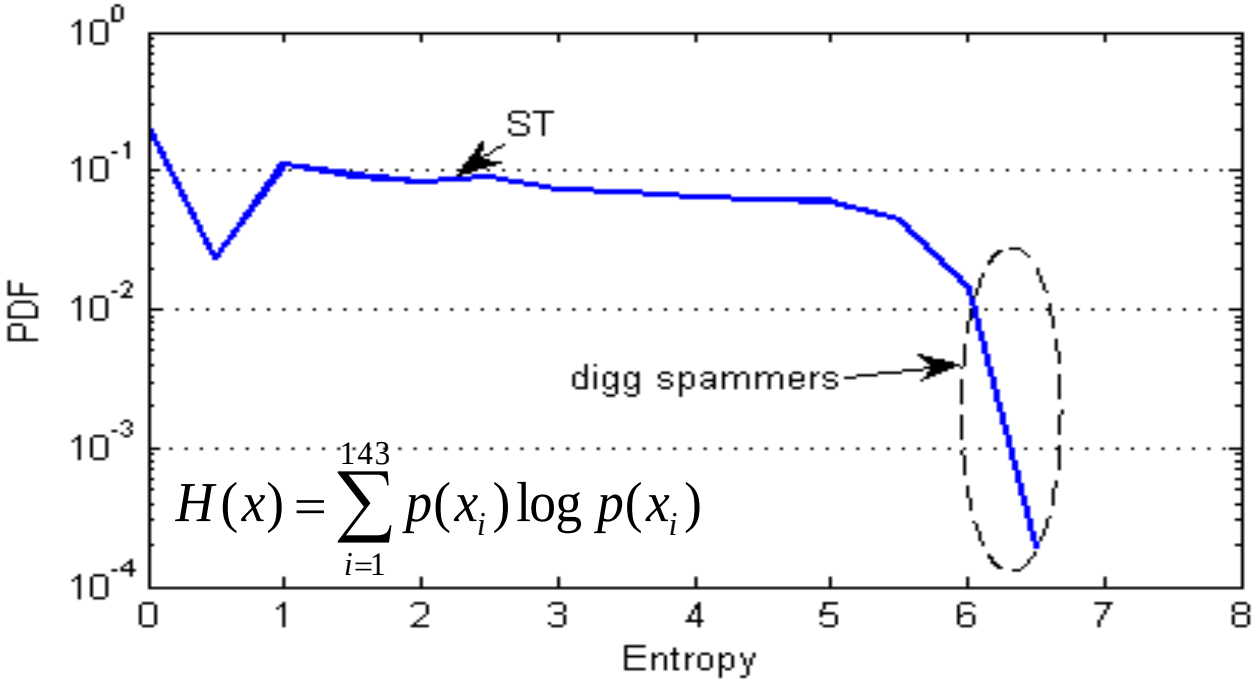
Diggs vs. inter-digg time intervals





1. Over **35%** diggs submitted within 1 min following their previous diggs
2. Over **12.75%** diggs submitted within 5 seconds following their previous diggs
3. Do spam diggs exist? (e.g., automatic scripts)

Entropy: measure randomness of a user's digg activities



- Inter-digg times split into 143 bins, by sec, min, hours (1-24, > 24)
- Compute each user's entropy of diggs
- Evidence of spam diggs
  - E.g., *Subvert and Profit* charges advertisers for votes in Digg

Outline

- ~~Analyzing structural properties of Digg social network~~
- ~~Measuring user digg activities~~
- Understand impact of social network on user diggs, content promotion and content filtering

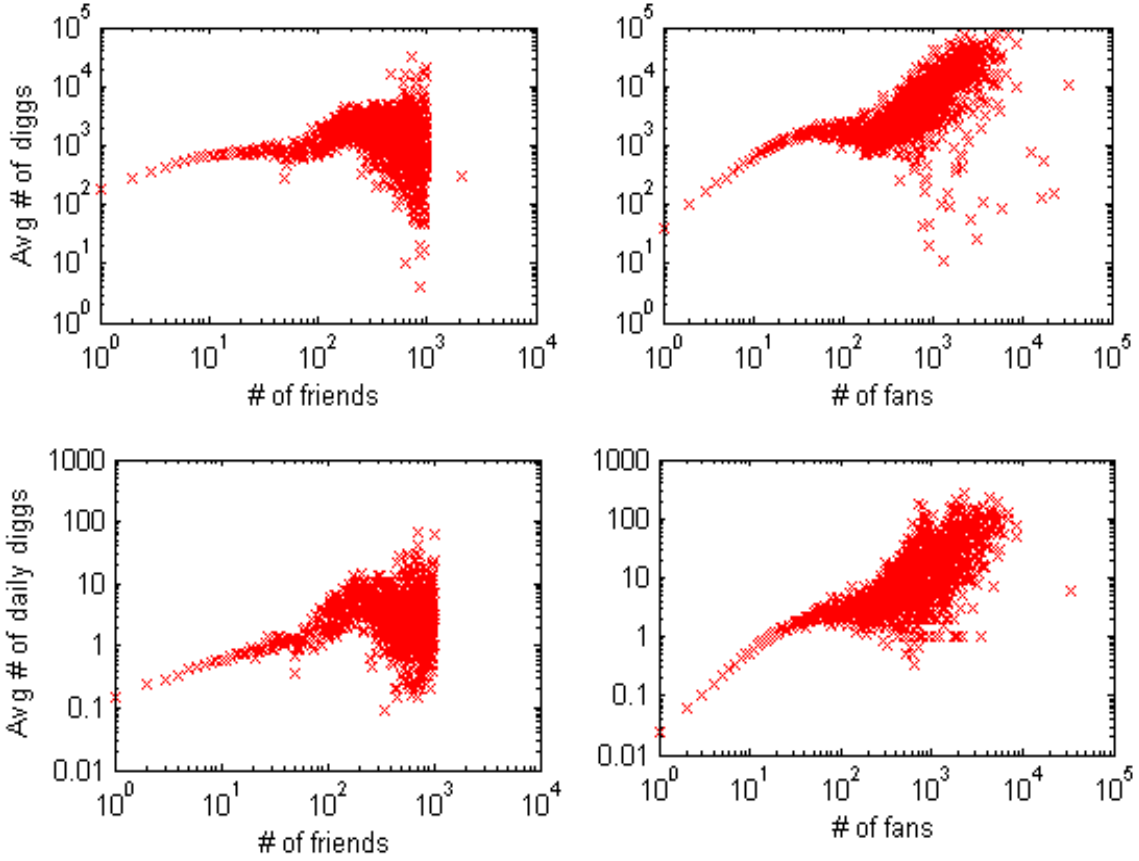
Impact of social links on user diggs

- Want to answer two questions:
  - Do people digg more actively if they have more friends?
  - Do people digg more actively if they are befriended by many others (celebrity pressure)?

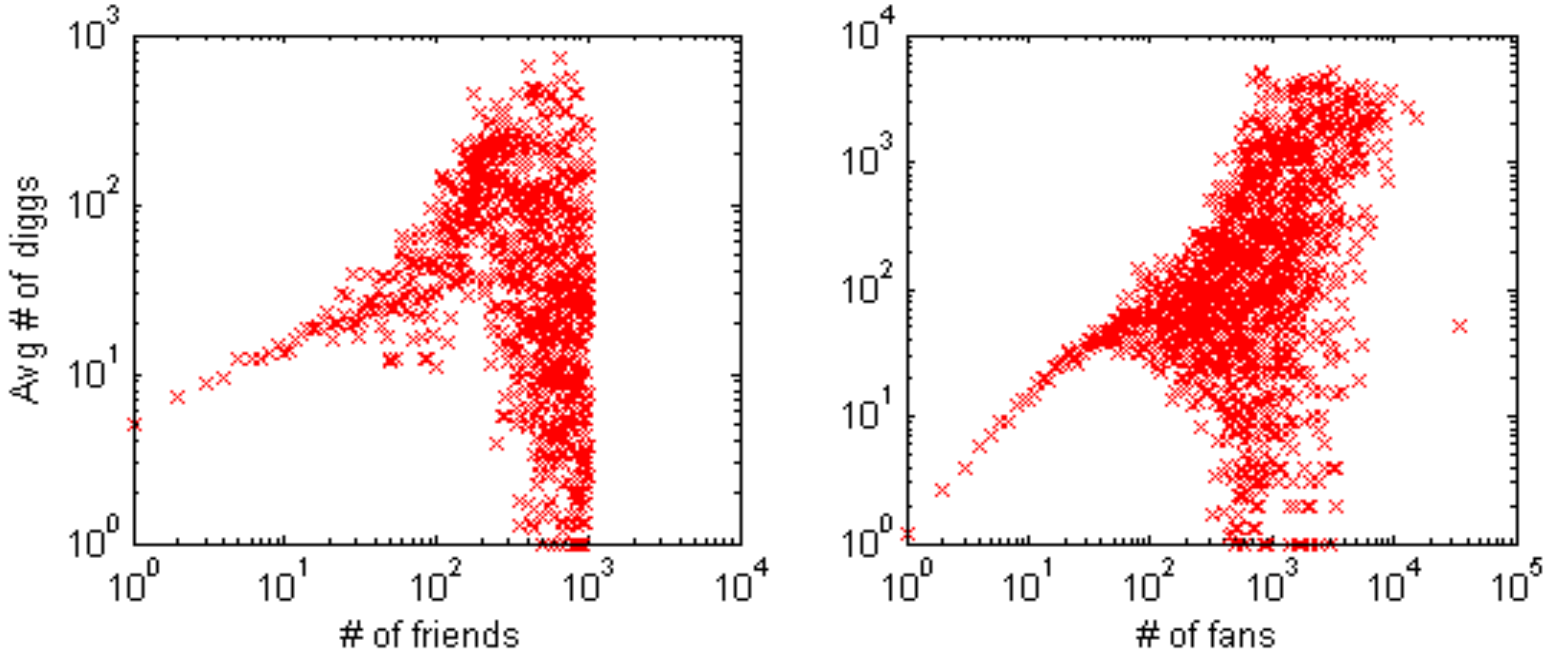
Diggs vs. social links in PT

Speculations on Diggs vs. fan links:

1. Higher visibility by more diggs, thus attracting more fans
2. Respond to celebrity pressure
3. Users with more fan links has been in system longer (older age), accumulating more diggs

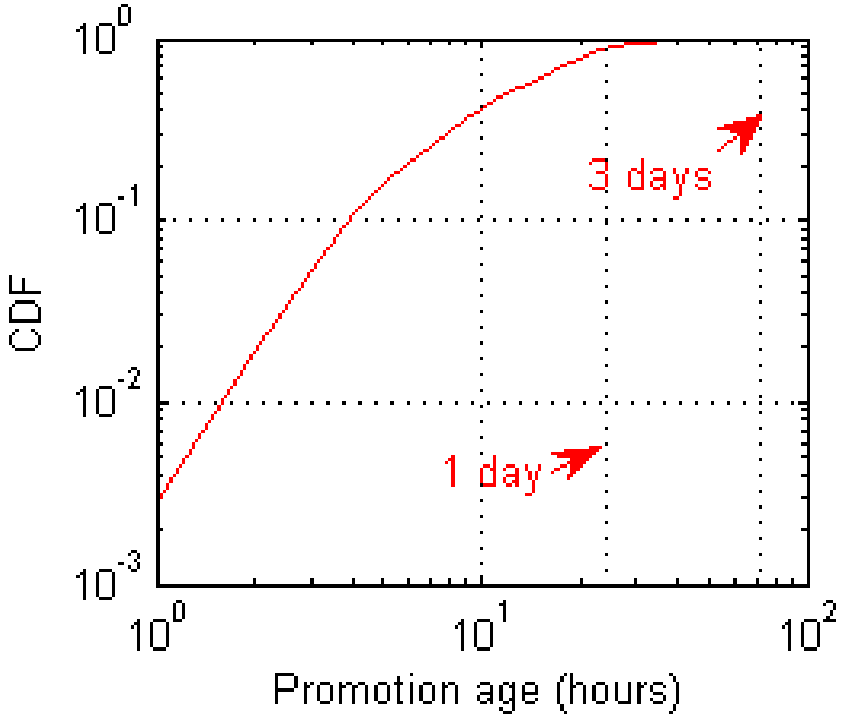


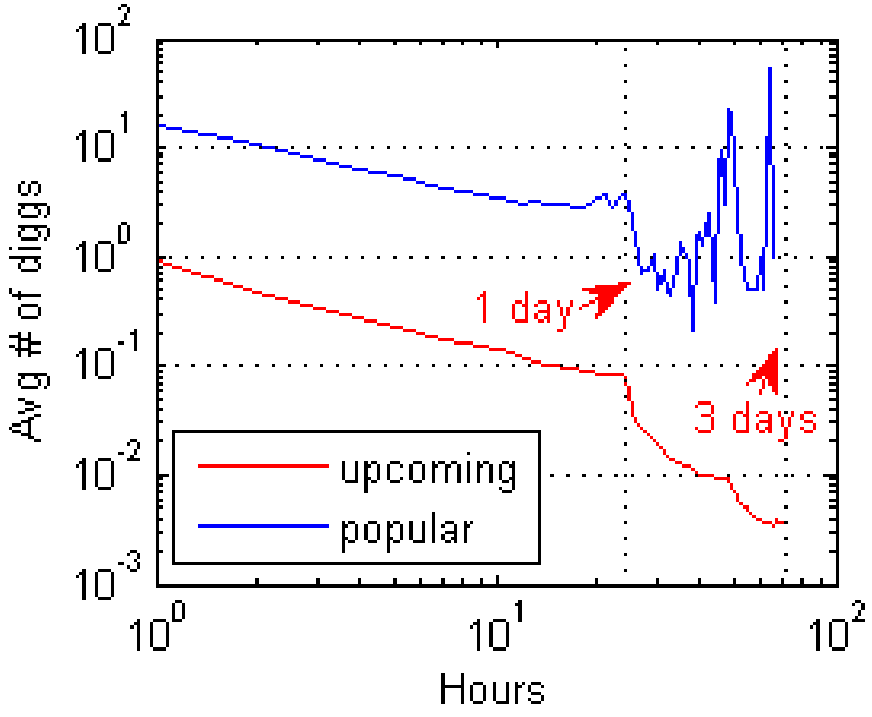
Diggs vs. social links in ST



- ST minimizes impact of user's age on the correlation
- The same observations hold in ST

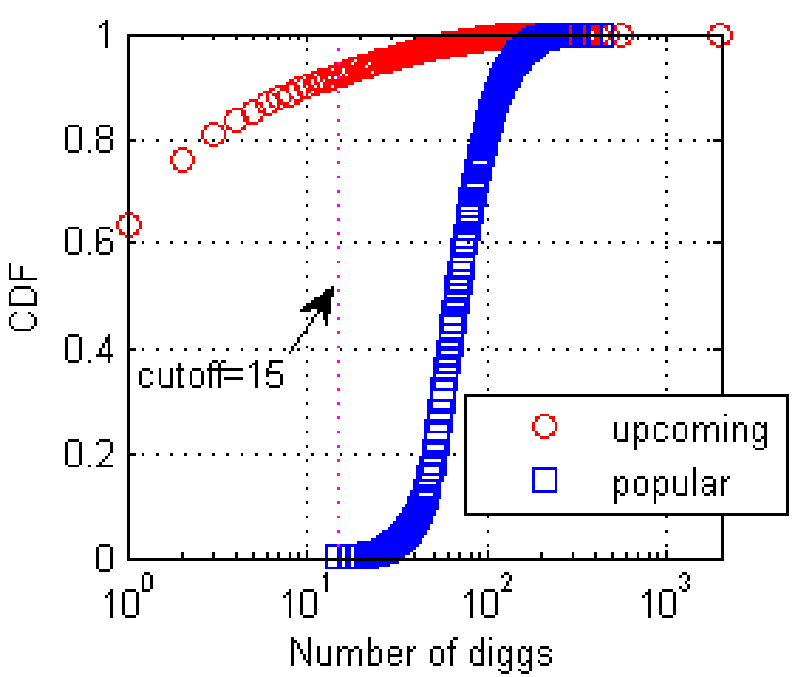
Content promotion

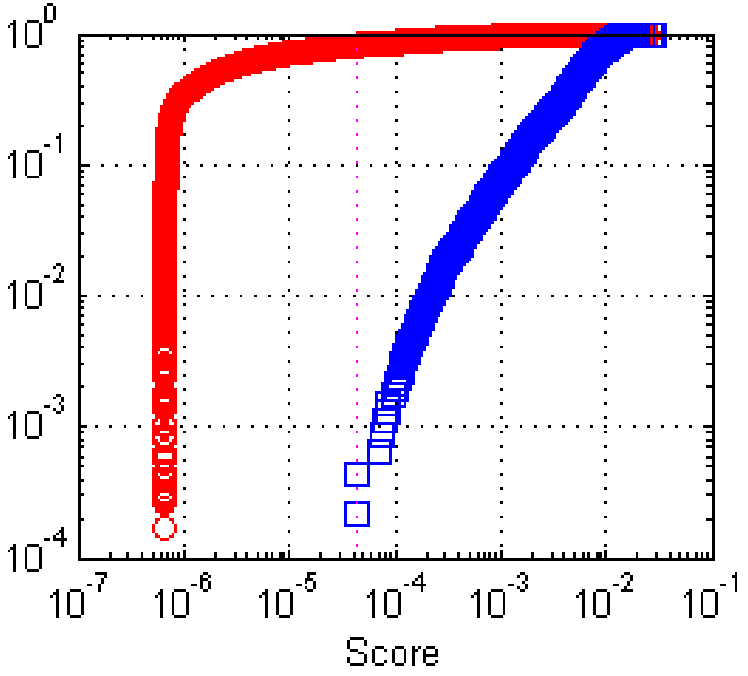




- Stories, if got promoted in ST, became popular within 3 days of their ages
- Stories, before promotion, received one order of magnitude higher digg rate than upcoming stories

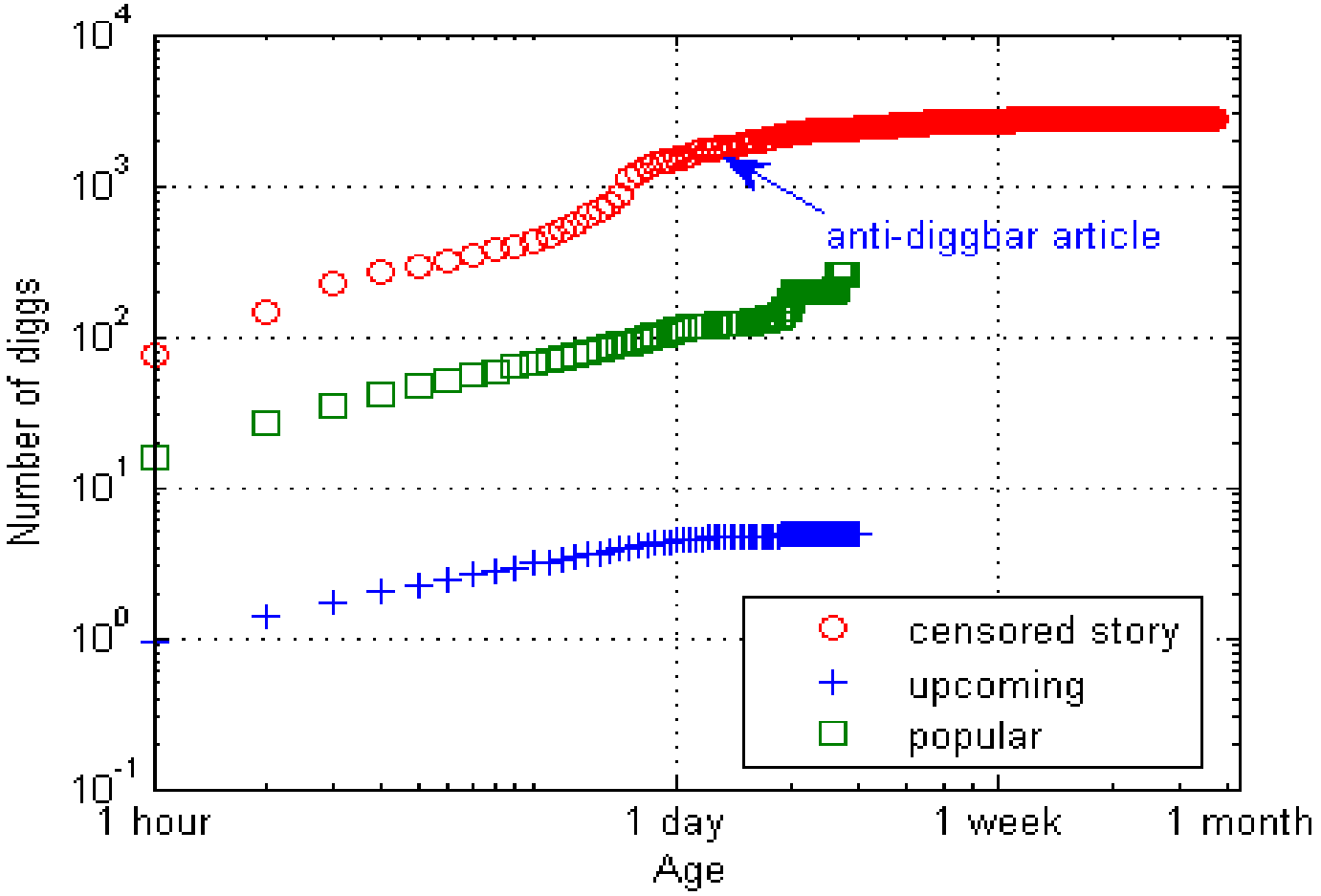
Content promotion: simple aggregation of diggs?





- # of received diggs is important to story promotion
- We speculate *Digg does not treat each digg equally*
  - Exploit PageRank and low link symmetry to weight individual diggs
  - 7.9% of upcoming stories received same or higher diggs, but subsumed in PageRank score.

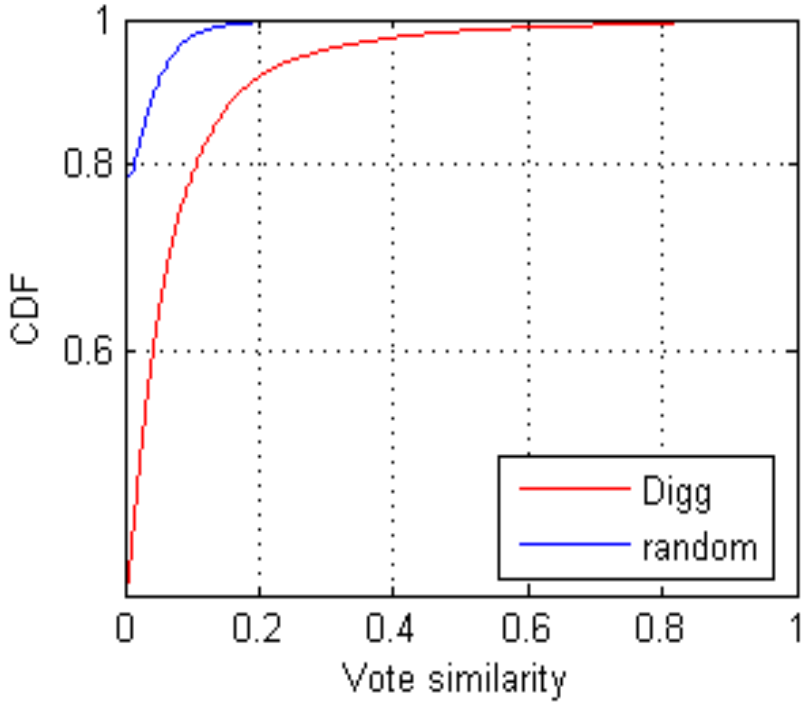
Content promotion vs. censorship

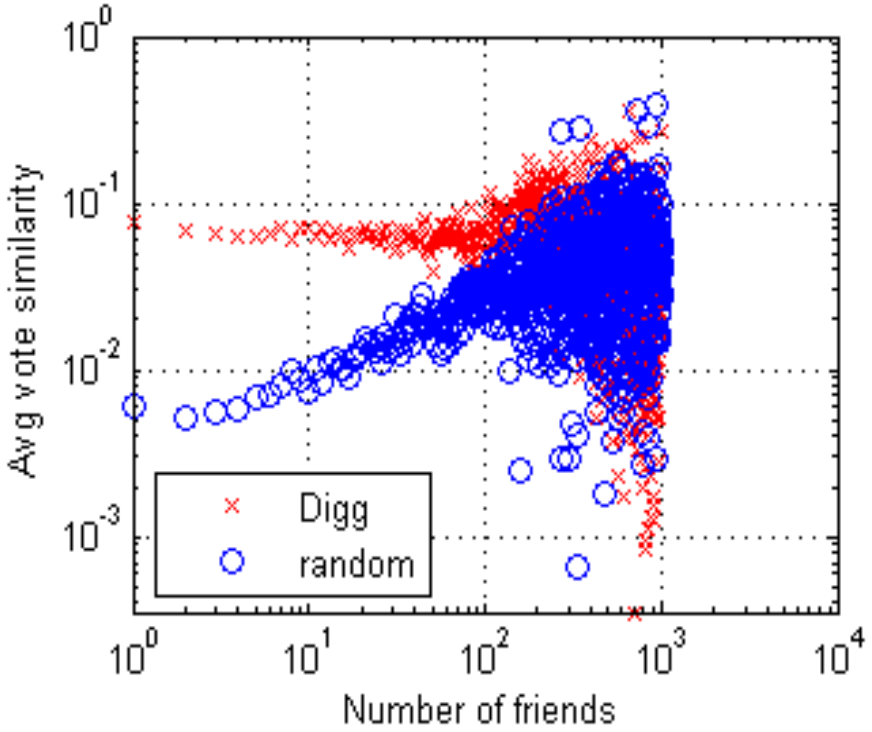


Content filtering

- Presents interesting content to users
- Influences users viewing and rating content
- Two filters in Digg
  - The friends interface
  - The front page

Content filtering vs. friends interface





- Vote similarity is computed between each user and her friends, using VSM
- The friends interface influences users with a small number of friends (<= 200)
- May need a better recommendation interface to present interesting content

Content filtering: front page

	Upcoming stories	Popular stories
Diggs	-95.2%	455.9%
Comments	-94.1%	559.8%
Total	-95.1%	462.2%

Popular stories: assume promotion age is t, then compare [0,t] and [t, 2t]
Upcoming stories: t = 72 hours

- The front page significantly influences users viewing and rating content

Summary

- Showed Digg social network differs from other previously studies OSNs
- Explored impact of social links on user diggs
  - Indicated spam diggs
- Examined content promotion
  - Provided evidence of content censorship
  - Showed presence of influential users (in the paper)
- Assessed content filtering
  - The Friend interface
  - The front page (content promotion)


Q&A

Thank You!