

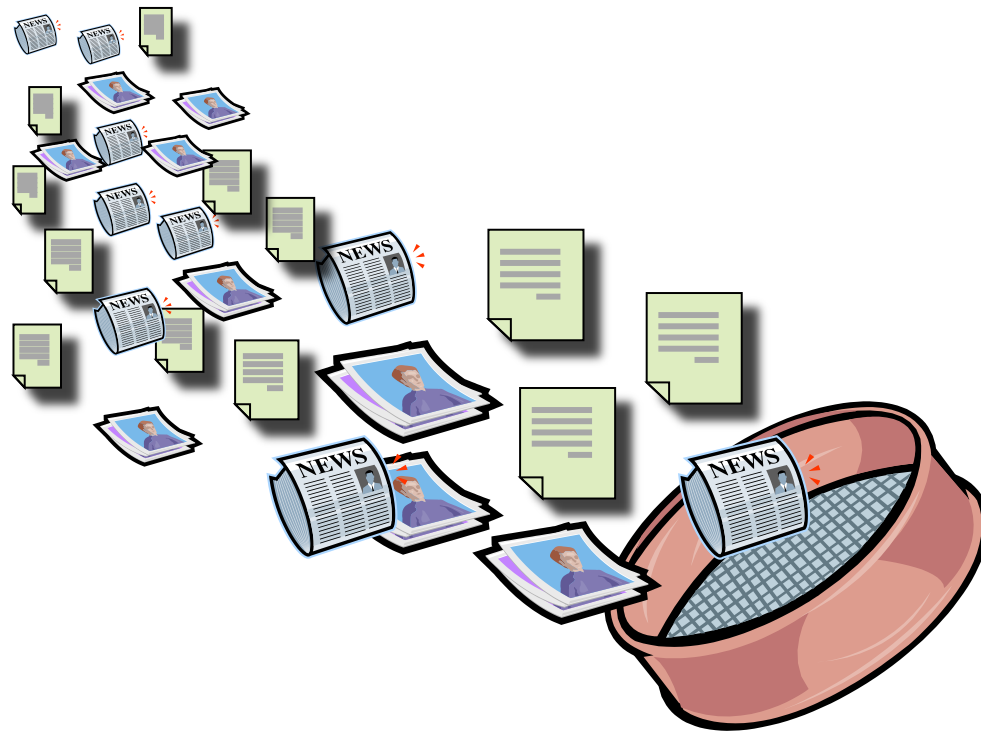
Using a Model of Social Dynamics to Predict Popularity of News

Kristina Lerman
USC Information Sciences Institute
lerman@isi.edu

Tad Hogg
Institute for Molecular
Manufacturing
taghogg@yahoo.com

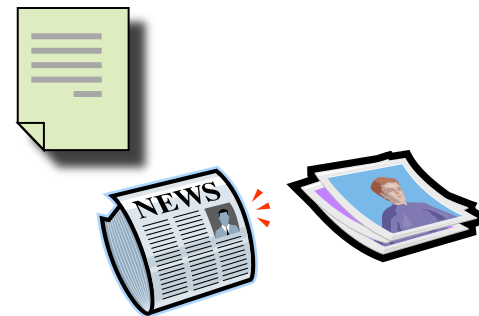
A large, stylized, light-colored 'V' logo is positioned in the bottom right corner of the slide, partially overlapping the text for Tad Hogg.

Explosion of content in social media



Daily growth of user-generated content

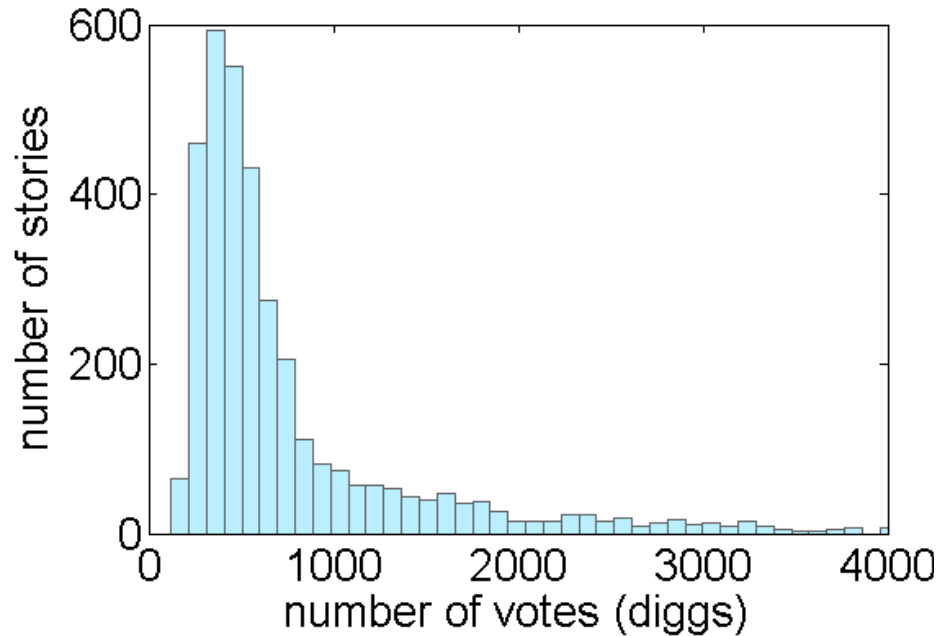
- 1,000,000+ new blogs posts
- 500,000+ new videos on YouTube
- 300,000+ new images on Flickr
- 16,000+ new stories on Digg
- ...



Goal: identify content that will become highly popular

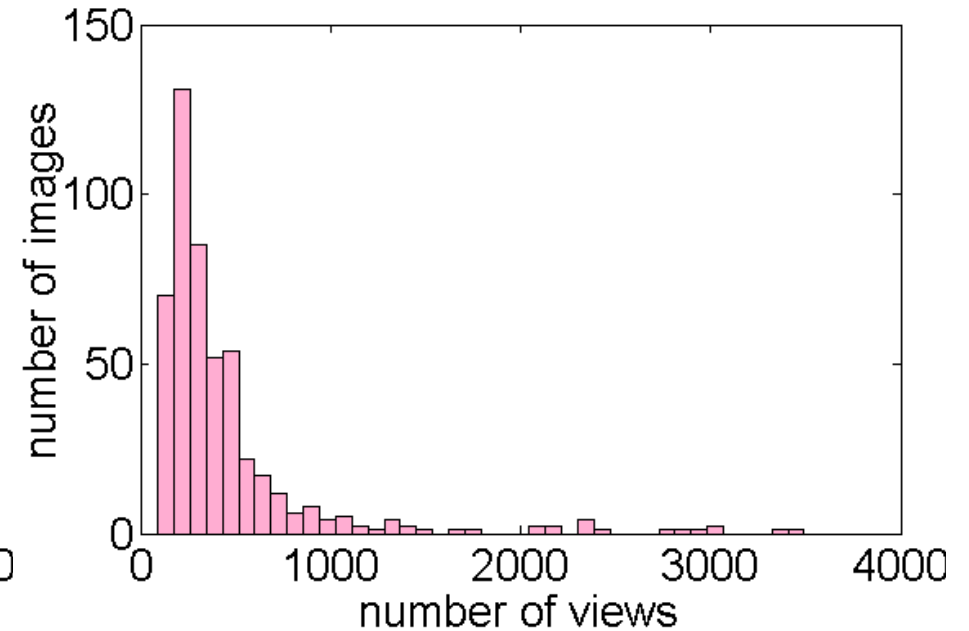
Inequality of popularity

Digg stories



Distribution of votes received by 3500 recent Digg stories in June, 2009

Flickr images



Distribution of views received by 500 images featured on Flickr's Explore page over a week in June 2006

Unpredictability of popularity

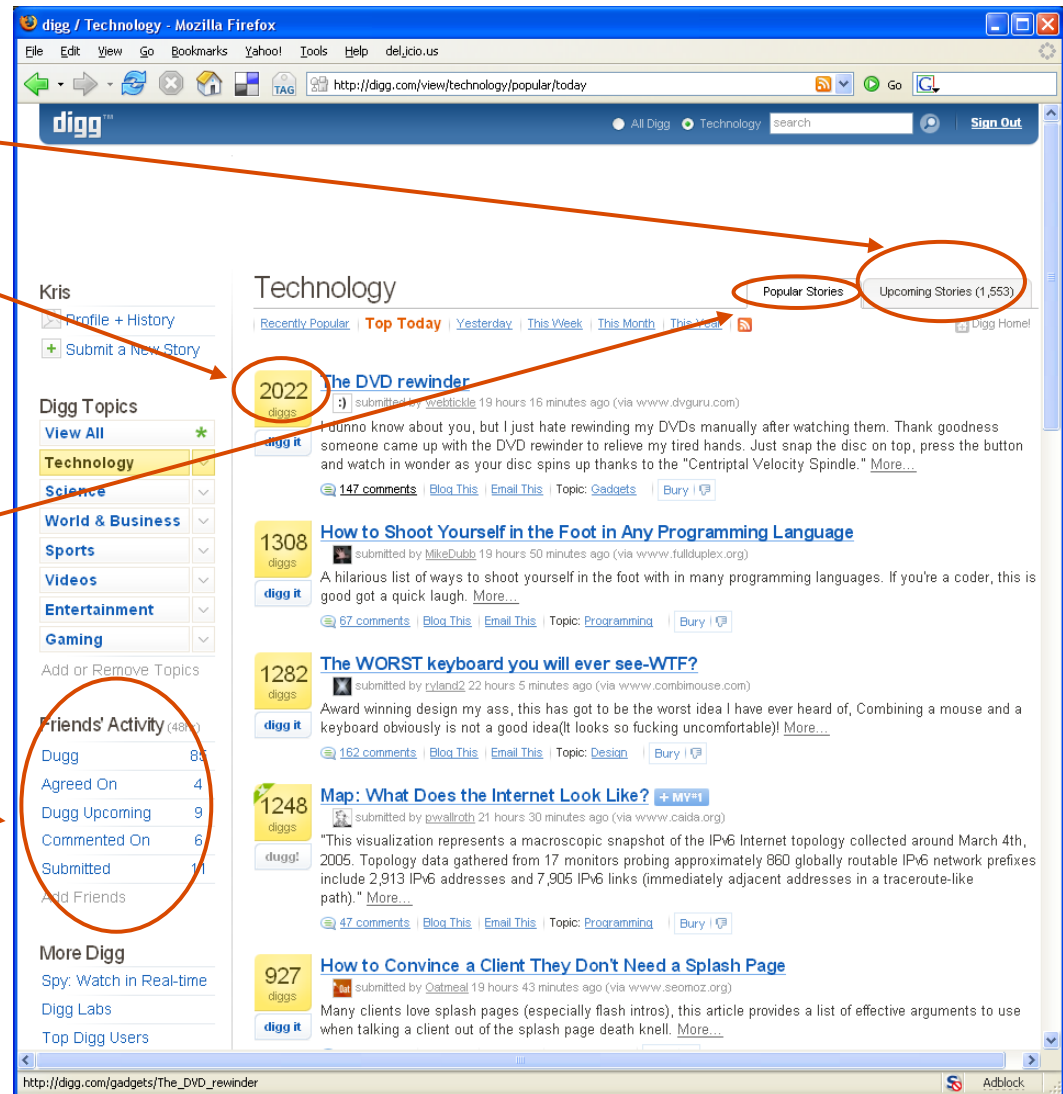
- Popularity is difficult to predict, even to experts
 - Quality – inherent feature of content
 - Social influence
 - i.e., knowing about the choices of others
 - Other factors?
- Artificial cultural market experiments [Salganik, Dodds & Watts, 2006]
 - Studied popularity of cultural artifacts, like music
 - Quality contributes weakly to popularity
 - Social influence is responsible for both **inequality** and **unpredictability** of popularity

Models of social dynamics allow us to predict popularity in social media

- Model of social dynamics
 - Stochastic model describes user response to content
 - E.g., evolution of number of votes a story receives on Digg
 - Accounts for user interface and social influence
- Model-based prediction
 - Observe users' early response to content to estimate model parameters
 - Future users' response specified by the model
- Test case: social news aggregator Digg

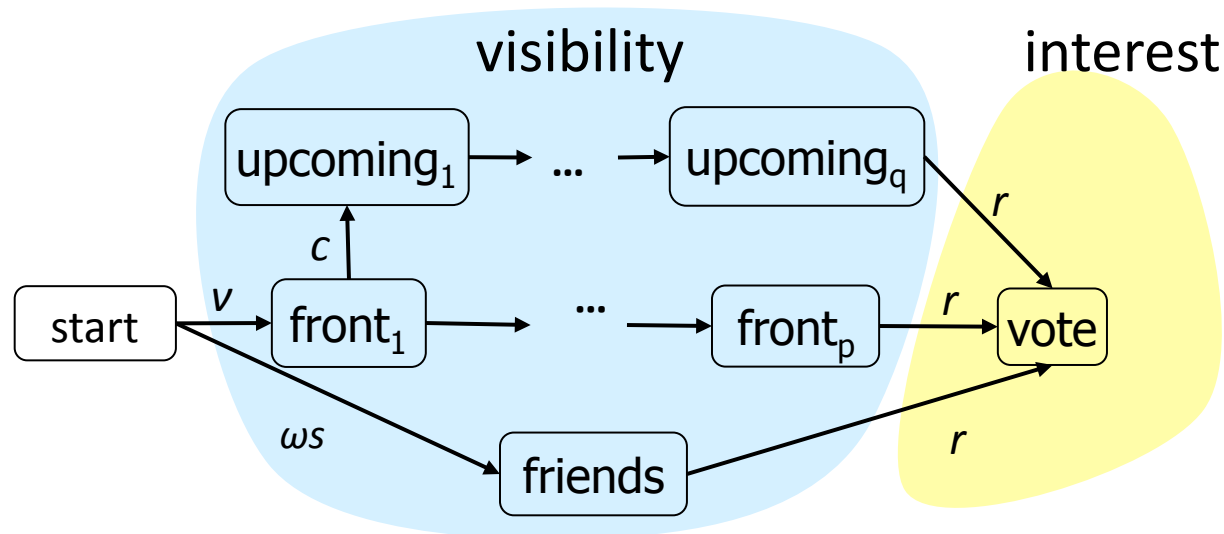
Lifecycle of a story on Digg

- user submits story to **Upcoming Stories**
- others vote on the story
- if story gets many votes quickly
→ promoted to **Front page**
- **Friends Interface** shows stories friends
 - submitted
 - voted on
 - ...



Stochastic user model

- visibility: does user **see** the story?
 - user interface
 - browse upcoming stories
 - browse front page stories
 - recommended by friends
- interest: does user **like** the story?



Collective model: number of votes for a story

$$\frac{dN_{\text{vote}}(t)}{dt} = r \overbrace{(\nu_f(t) + \nu_u(t) + \nu_{\text{friends}}(t))}^{\text{visibility}}$$

rate users find story

- on front page list: ν_f
- on upcoming list: ν_u
- via friends: ν_{friends}

fraction of users seeing the story who vote for it: r

Estimating model parameters



- parameters for
 - story visibility
 - story interestingness
- estimate from sample of users & stories

Digg data set

- votes vs. time for stories
 - for several days in May & June 2006
 - 2152 stories with at least 4 votes submitted by 1212 distinct users

510 of these stories promoted to front page
- number of fans for active users
- story rates:
 - submitted: ~ 60/hour
 - promoted to front page: ~ 3/hour

Story visibility

- user viewing behavior not available:
 - which stories users look at
 - how they find stories
 - front page, friends interface, ...
- estimate indirectly from models & data

Modeling story visibility



- story location
- navigating web sites
- number of fans

modeling story visibility



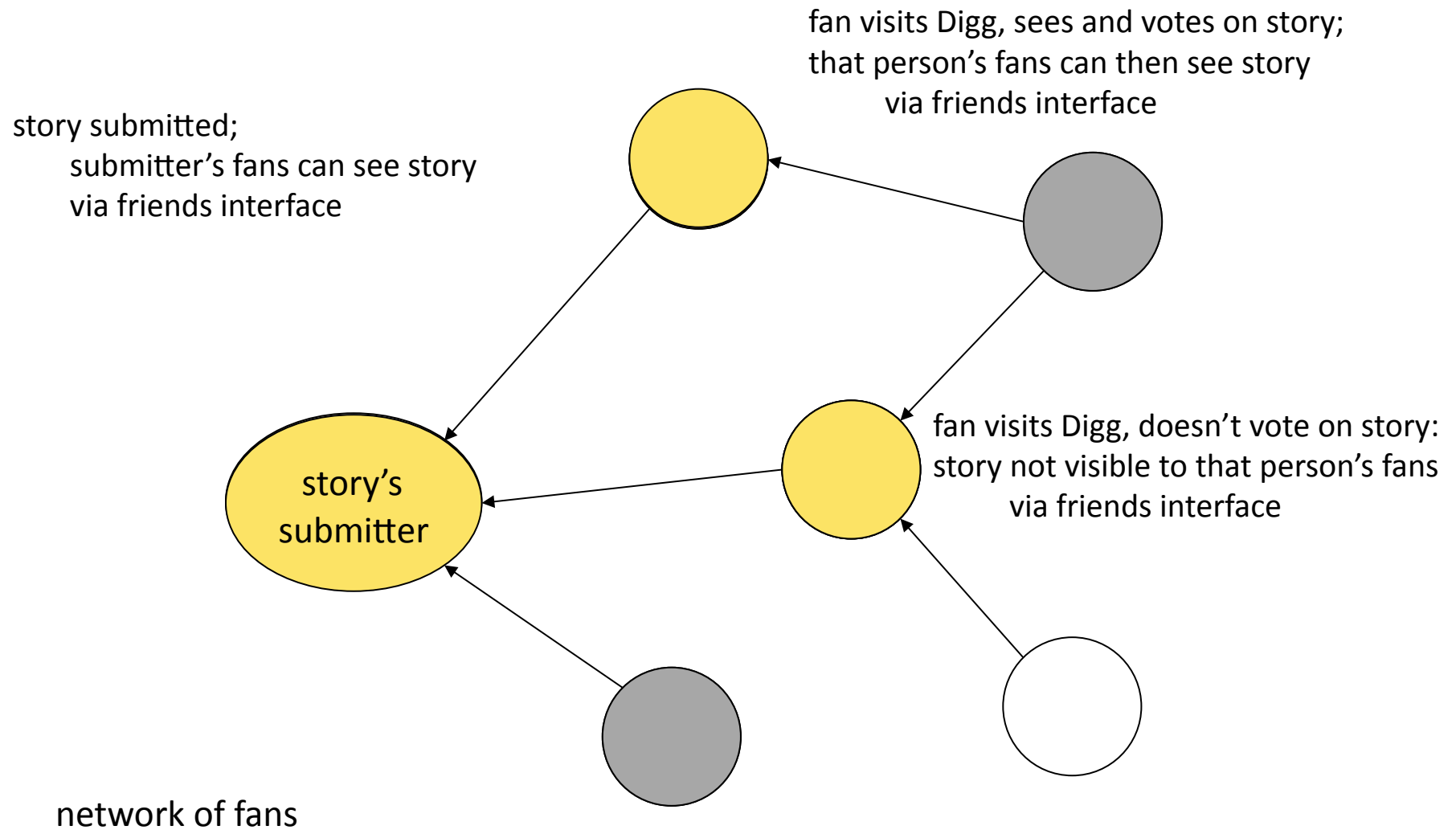
- story location
 - position on upcoming or front page
 - promotion to front page
- navigating web sites
 - “law of surfing”
 - Huberman et al., “Strong Regularities in World Wide Web Surfing”, *Science* 1998
- number of fans

Story location

Digg presents stories in lists

- 15 stories per page, most recent first
- a given story
 - moves down the list as new stories added
 - eventually moves to later pages
 - if promoted:
 - moves from *upcoming* to top of *front page* list

Story visibility via friends interface



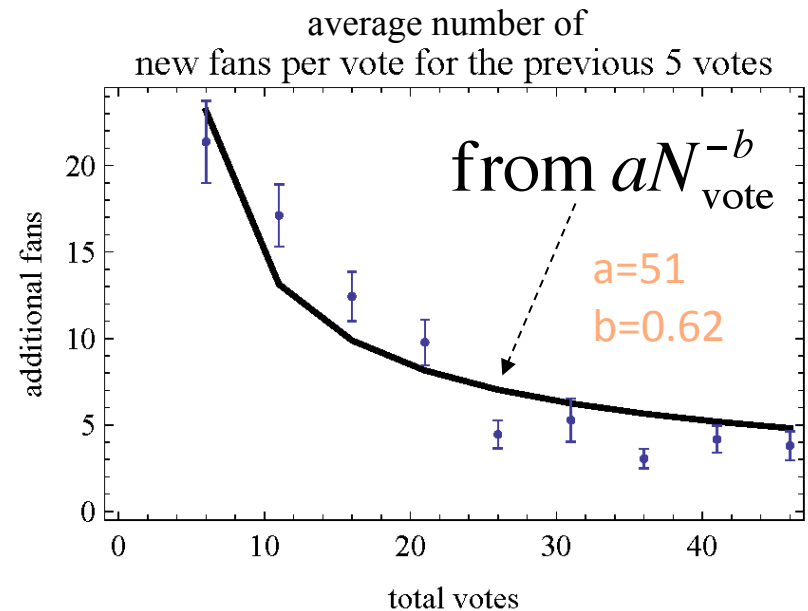
Story visibility via friends

- each voter enables her fans to see story
- number of fans not yet viewing story: $s(t)$
 - based on number of votes on the story
 - submitter's fans: $s(0)$

$$\frac{ds}{dt} = -\omega s + aN_{\text{vote}}^{-b} \frac{dN_{\text{vote}}}{dt}$$

fans of prior voters visit Digg

new fans from new votes



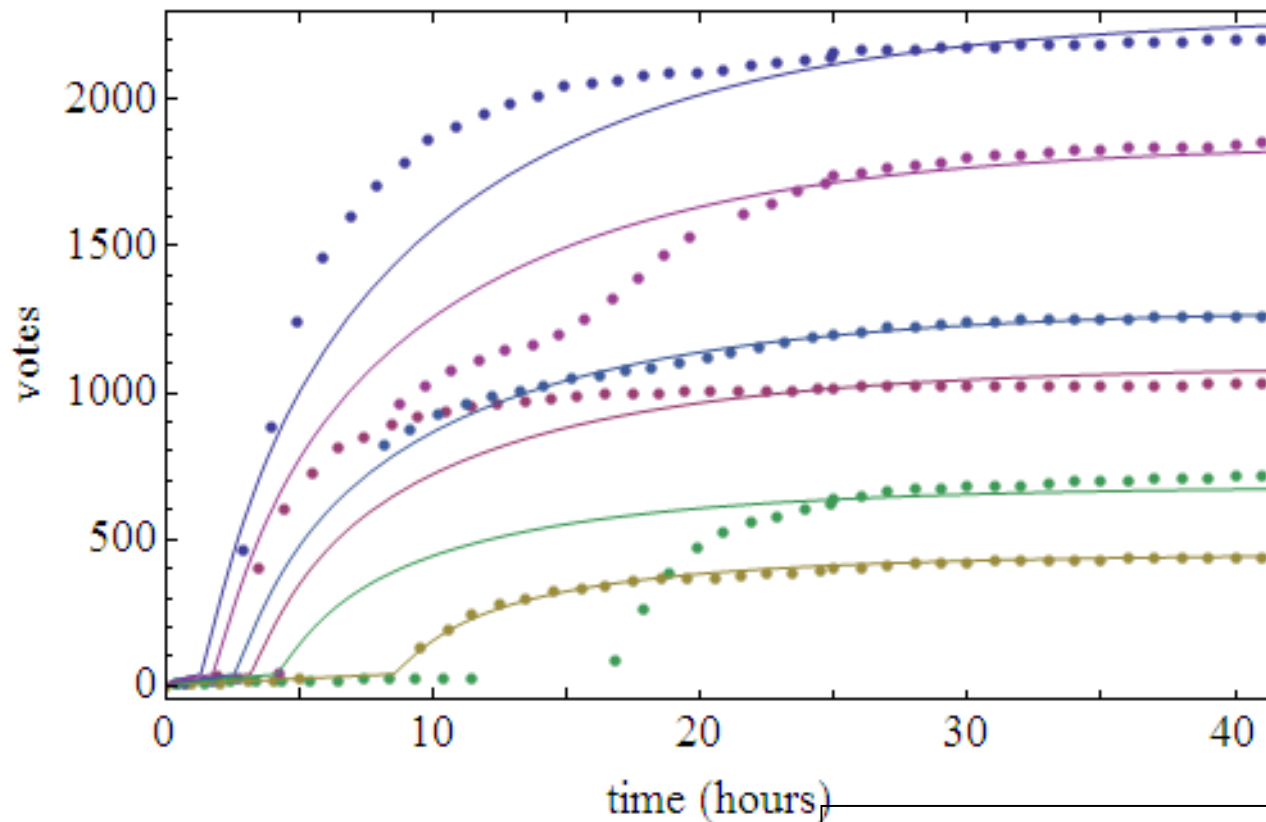
Story interestingness



- reasons users vote are not available, e.g.,
 - novelty [Wu & Huberman 2007]
 - popularity (determining interest, not just visibility)
 - e.g., “cool” fashion or gadgets
 - can test via web experiment [Salganik et al. 2006]
- estimate from models & data
 - fitting model to vote history after accounting for visibility

Votes vs. time

model vs. observations for 6 stories
(one parameter fit for r)



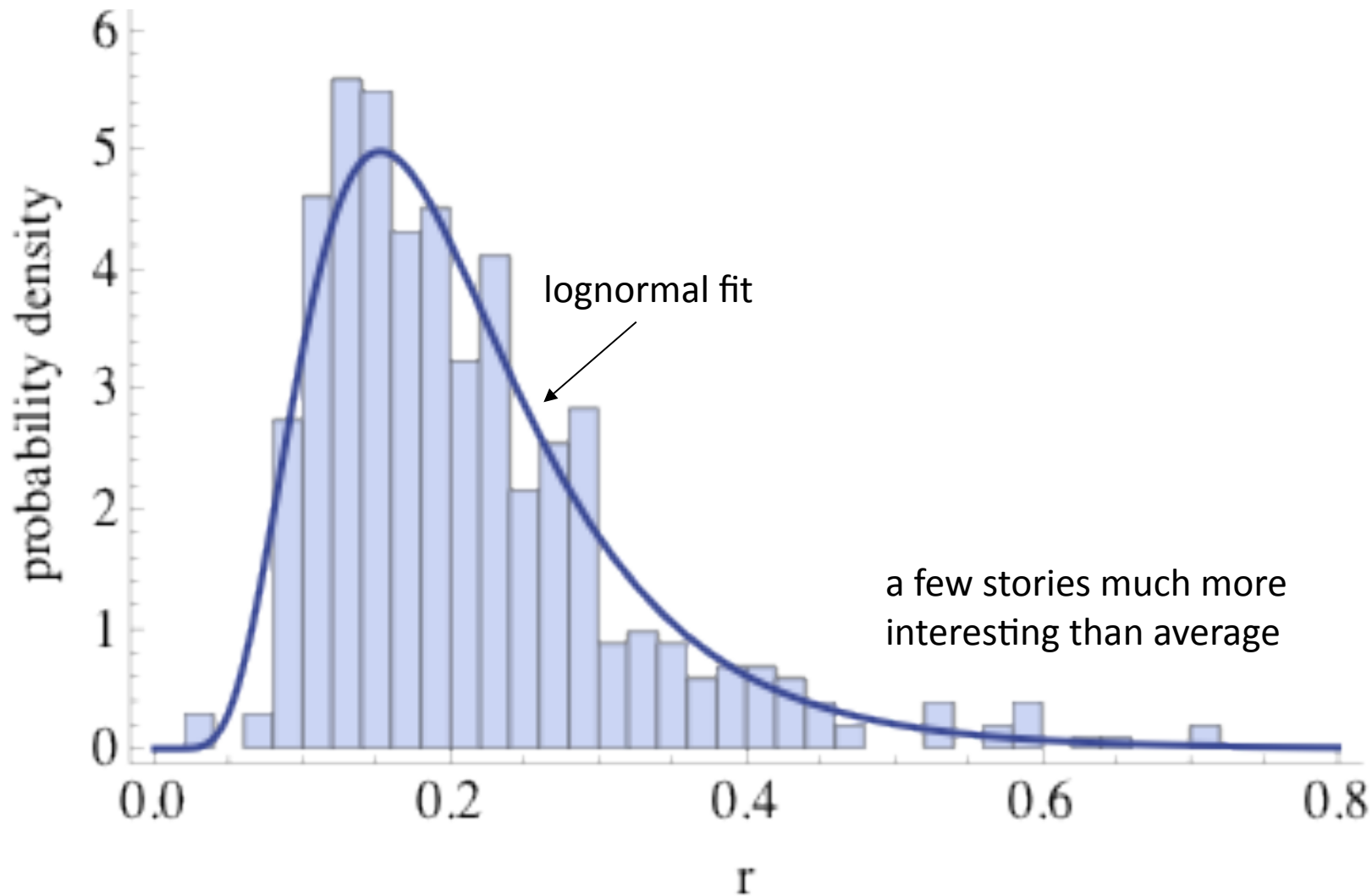
model captures qualitative features

- slow growth on upcoming pages
- influence of fans on promotion
- rapid growth if story promoted

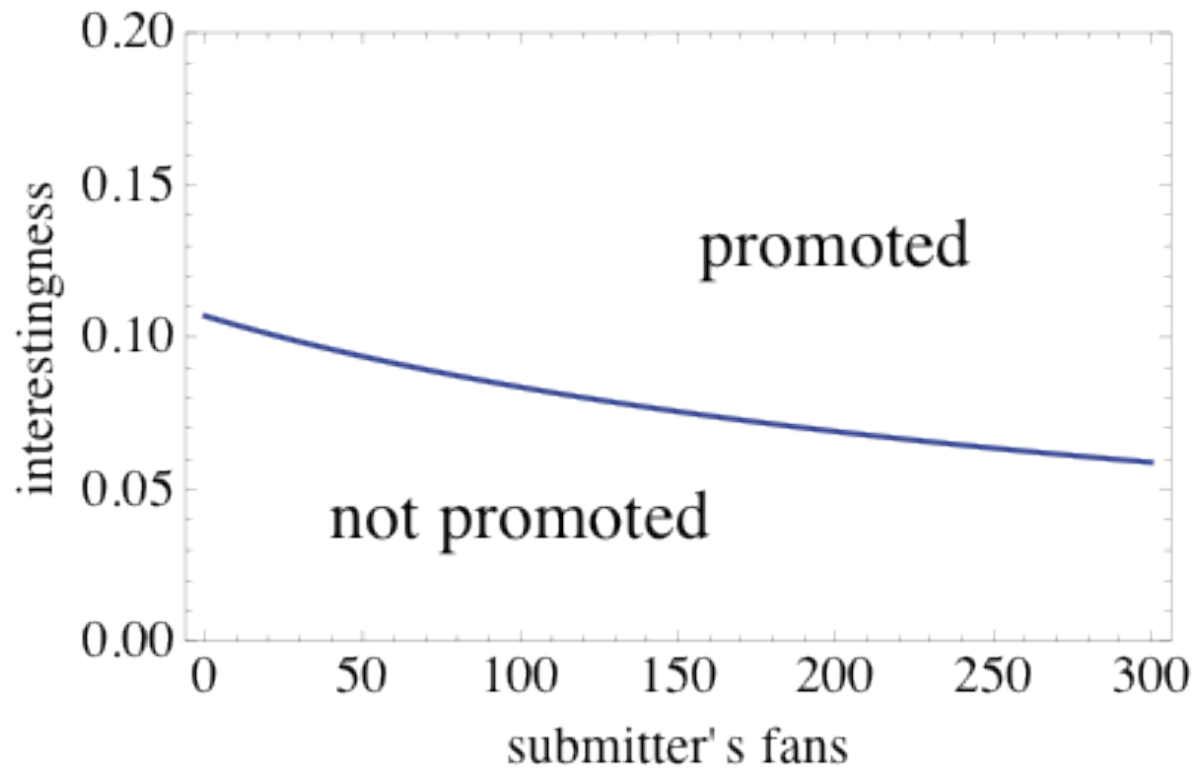
A model of social dynamics allows us to separate effects of story quality and social influence and study how each affects popularity

- Estimate story quality – interestingness r
 - “inherently unpredictable” [Salganik et al., 2006]
- Predict promotion
- Predict long-term popularity from early user reactions

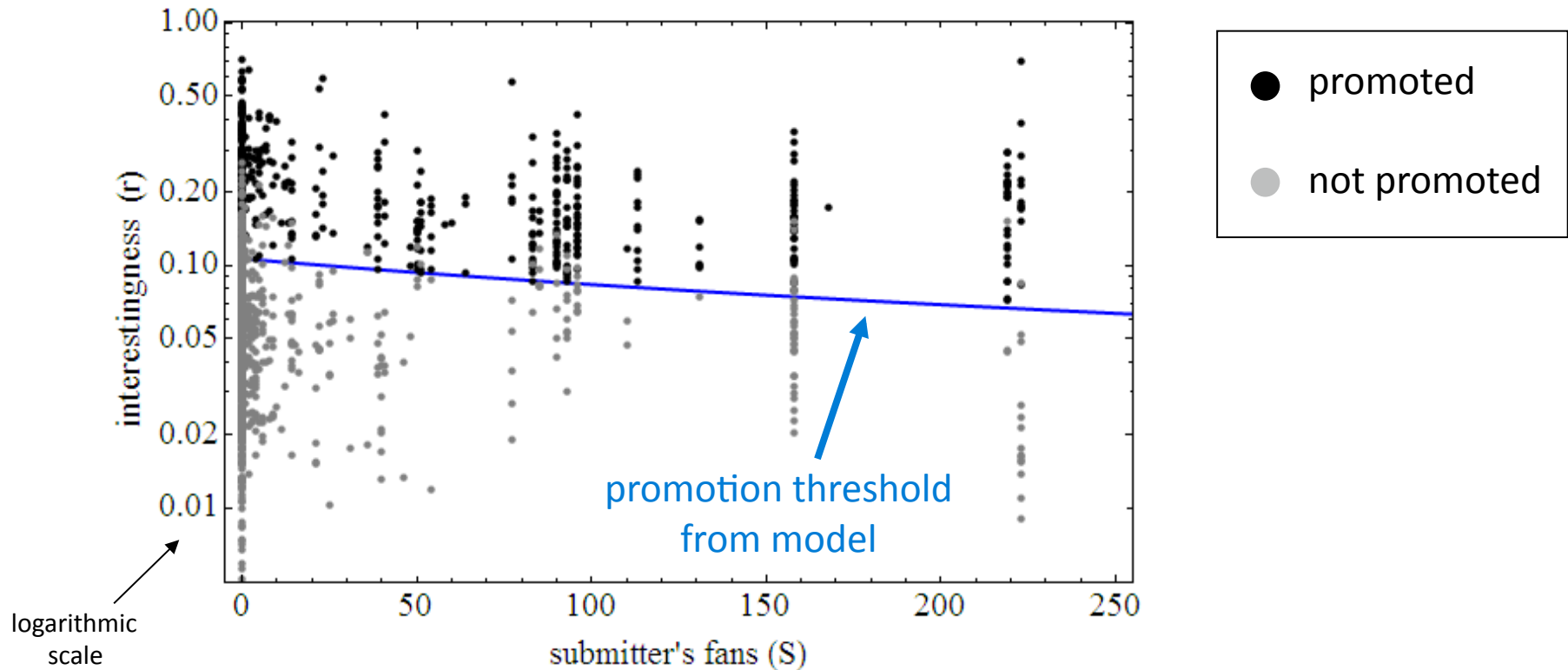
Interestingness for promoted stories



model: promotion to front page



Promotion prediction vs. data: 95% accurate



model prediction errors for promotion:
false negative: 1 out of 510
false positive: 108 out of 1642

Predict from early user reactions

- rationale: early reactions indicative of other users
- Model-based prediction
 - estimate story interestingness from early votes
 - use model to extrapolate to final votes
- accounts for variation in visibility
 - Due to the user interface
 - Due to social influence (friends interface)

example: prediction using model

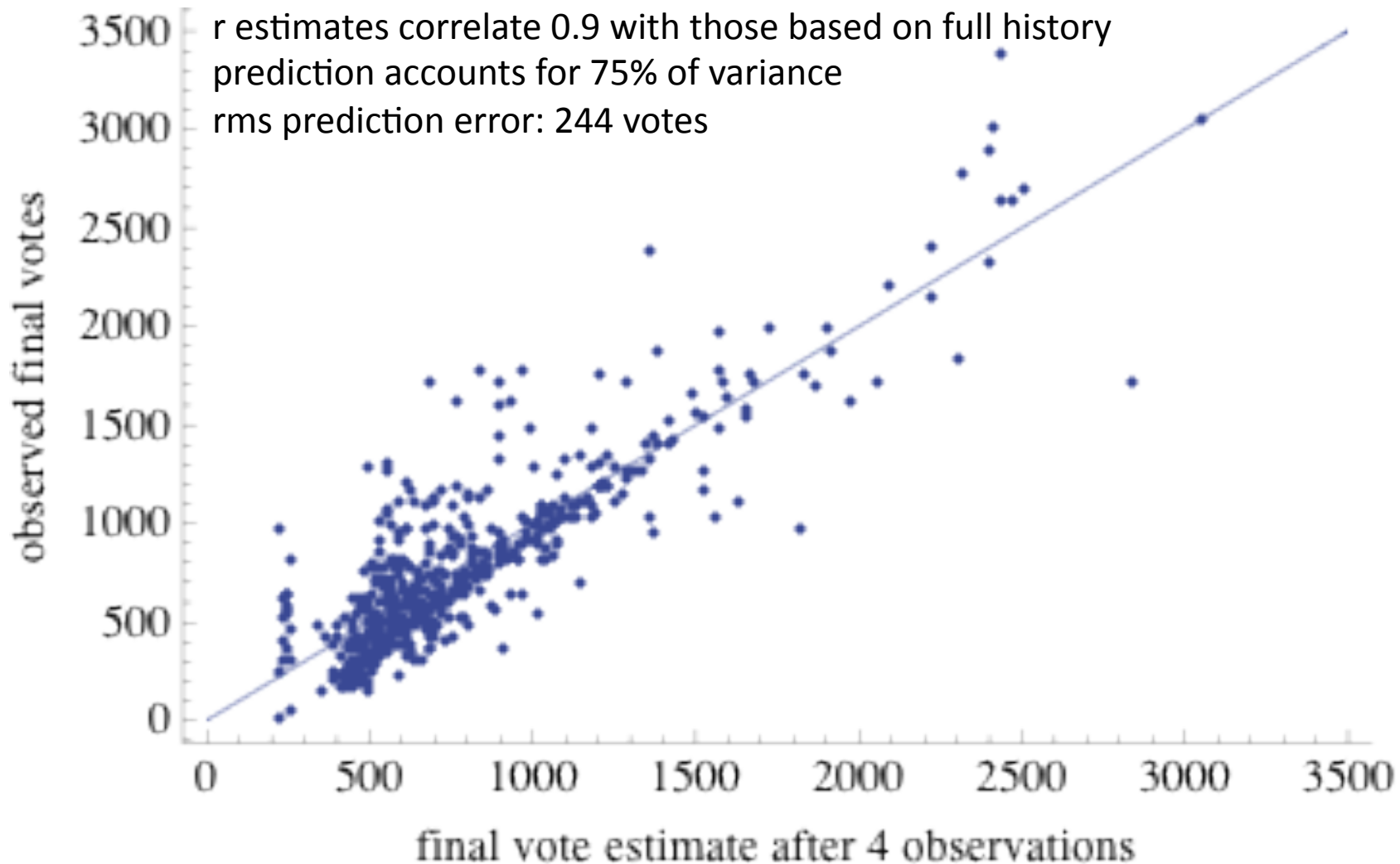


predict from first 4 observations

r estimates correlate 0.9 with those based on full history

prediction accounts for 75% of variance

rms prediction error: 244 votes



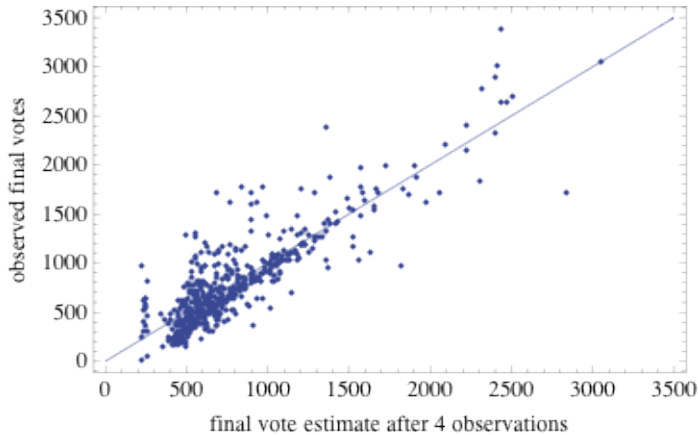


- Direct extrapolation from early votes
 - Useful for various communities, e.g., Digg & YouTube [Crane & Sornette 2008; Szabo & Huberman 2008]
 - Caveat: Applicable to promoted stories only
- Proportion of early fan votes [Lerman & Galstyan 2008]
 - Proportion of fan votes among early votes
 - Large fraction of early fan votes → not popular story
 - » Niche interest story?
 - Small fraction of early fan votes → popular story
 - » General interest story
 - Caveat: Applicable to stories submitted by well-connected top users

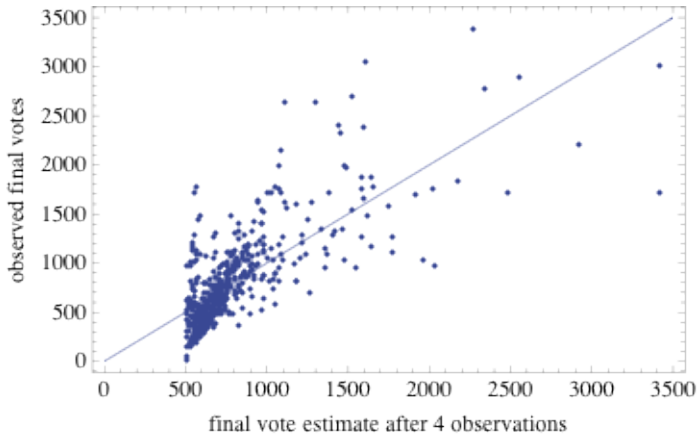
Comparison with “direct extrapolation”



full model



votes only



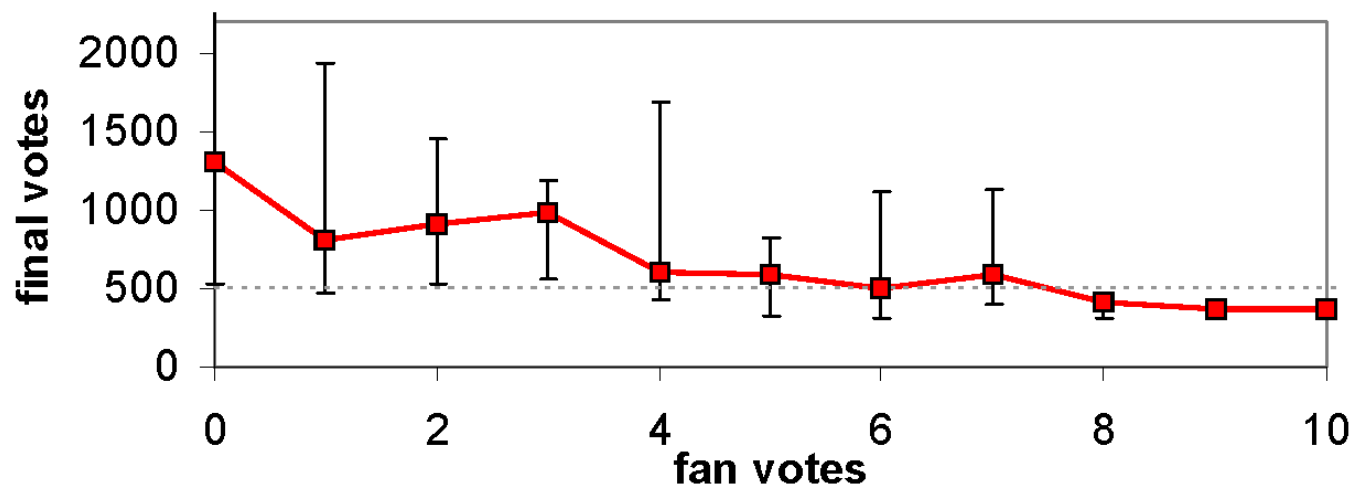
full model is better than not including visibility
(differences significant, p-value 10^{-4})

variance accounted for	75%	56%
rms prediction error	244	327

Prediction based on early fan votes

- Popular stories initially receive fewer fan votes
 - Train classifier to predict if story will receive > 500 votes
 - [Lerman & Galstyan, *WOSN* 2008]

Story popularity vs number of fan votes within the first 10 votes



Comparison

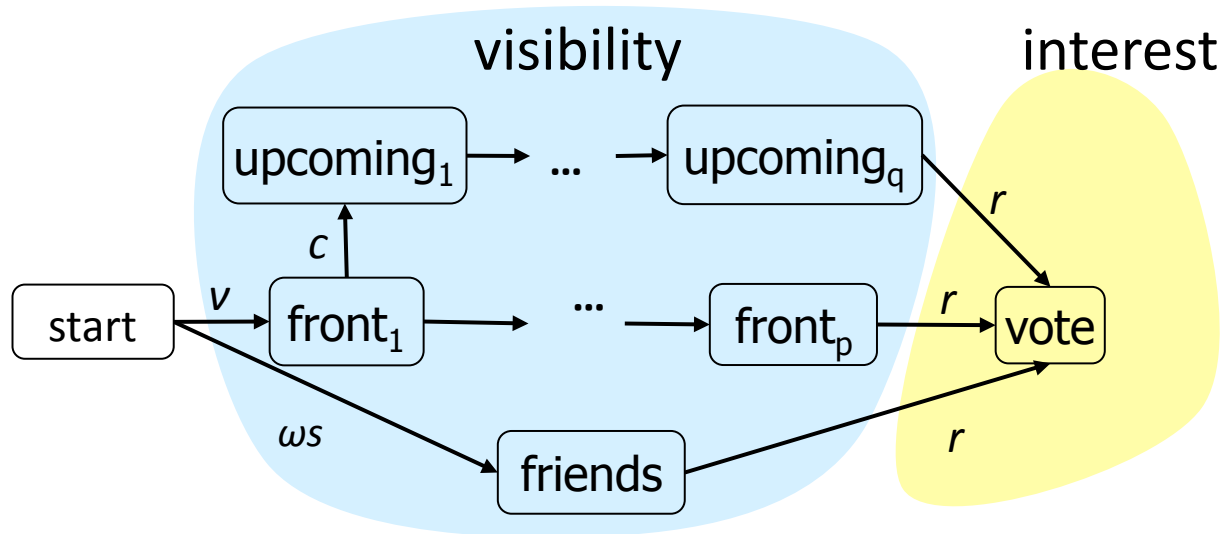
Predict whether 39 stories submitted by top users will become popular
(task: predict whether the story will receive more than 500 votes)

Prediction method	Early fan votes	Social dynamics model	Digg promotion
Number of votes used	First 10	First 10	First >40
Precision (all 39 stories)	0.21	0.27	-
Precision (14 promoted stories)	0.5	0.43	0.31

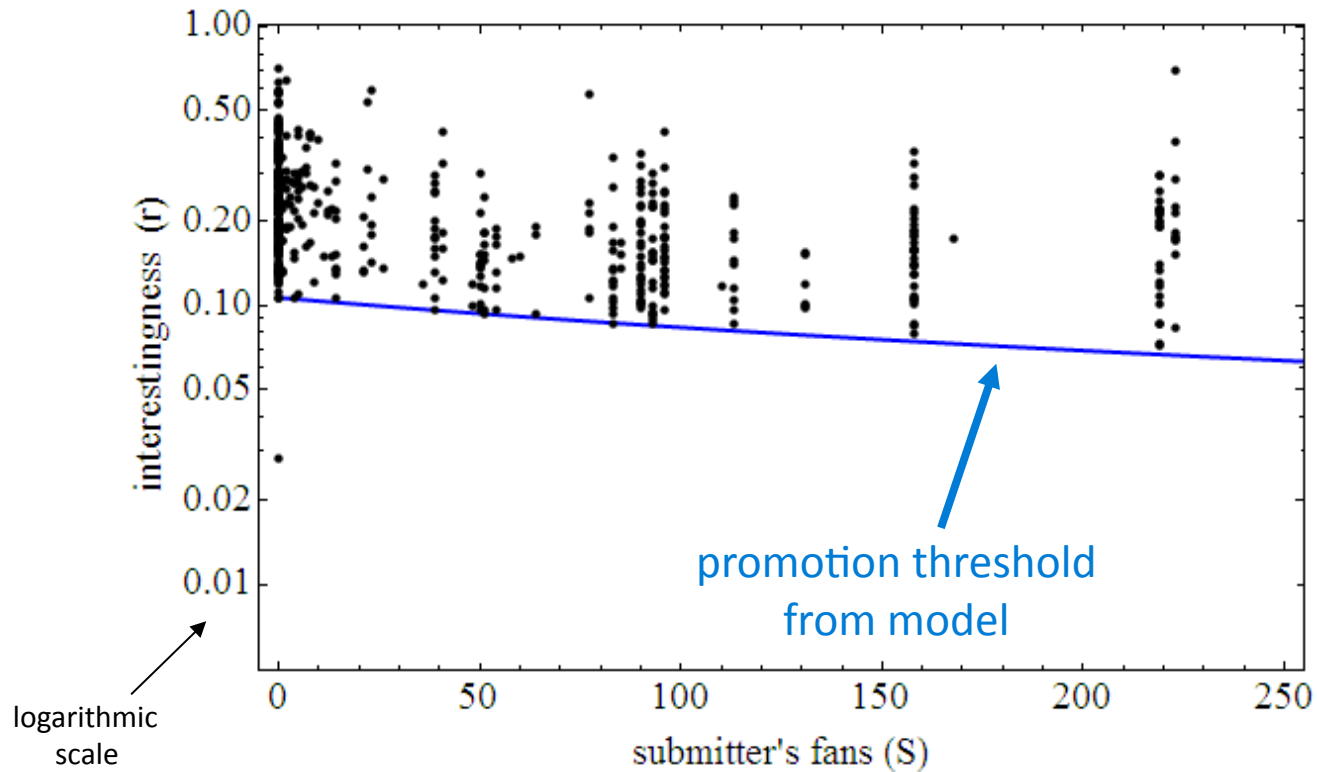
- The open nature of social media enables prediction
 - Observations of user behavior lead to model of social dynamics
 - Social dynamics models are a powerful tool to explore social media
 - Estimate unmeasurable parameters, e.g., story quality
 - Predict future behavior
- Social dynamics models
 - Extend the model to allow fans and non-fans to have different interests
 - Discover niche interests
 - Model user diversity

Stochastic user model

Stochastic user model based on Digg user interface

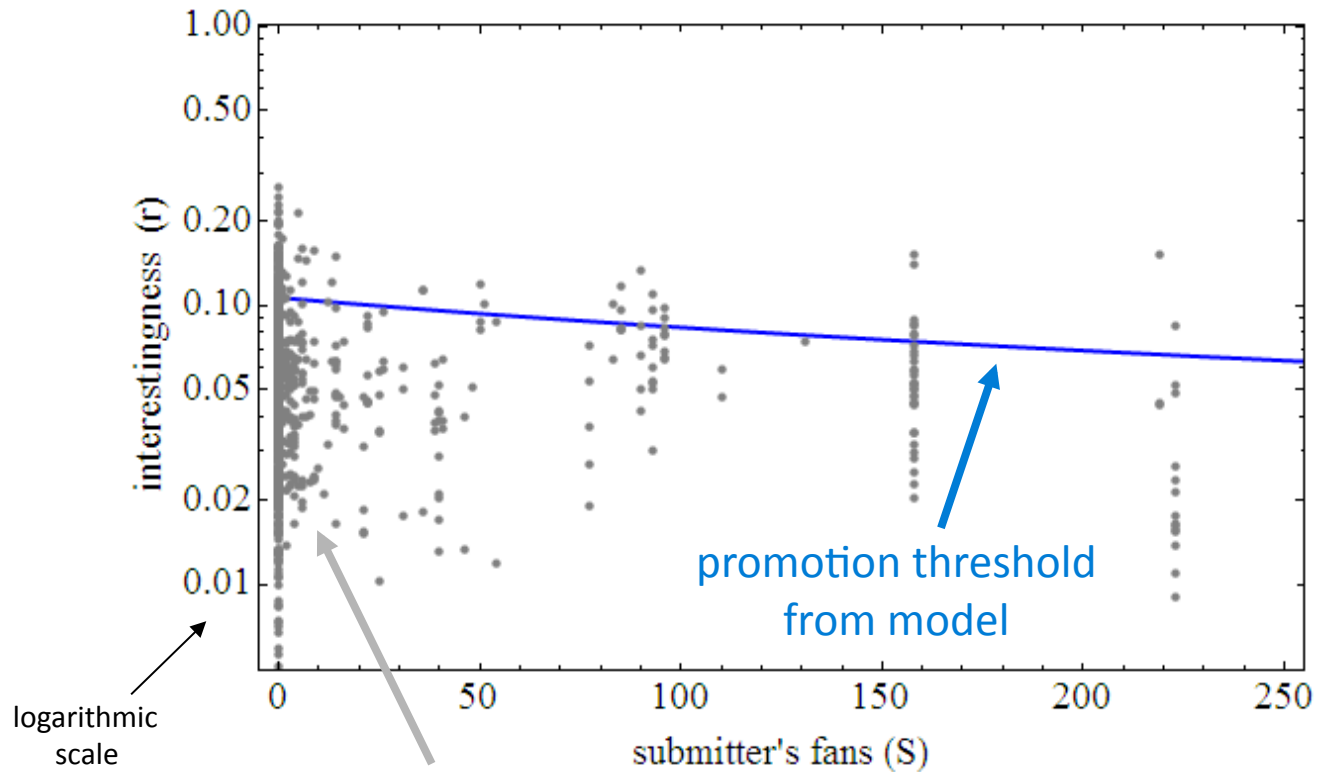


promotion prediction vs. data: promoted stories



model prediction errors for promotion:
false negative: 1 out of 510

promotion prediction vs. data: non-promoted stories



model prediction errors for promotion:

false positive: 108 out of 1642

votes	<i>r</i>	story title
3054	0.71	Lego Aircraft Carrier Complete!
3388	0.70	How to Make a Spider from 5 Crisp Dollar Bills (and Scare Waitresses!)
3125	0.65	Things You Didn't Know About Your Body
2981	0.63	25 Worst Tech products of all time
2776	0.59	The Coolest Solar Eclipse Photo You Will Ever See...
2748	0.59	14 year old kid becomes millionaire through online scamming
2701	0.58	X-Men: Last Stand Post-Credits Scene?
2327	0.58	18 Days of Reckless Computing
2690	0.58	First Photos of MIT's \$100 Laptop
1310	0.57	Nintendo Puts \$250 Price Tag on Wii OFFICIAL
2204	0.54	MacBook vent blocked
2413	0.54	Wii will cost less than \$220
397	0.09	Microsoft: "OpenDocument is Too Slow"
364	0.09	AMD aims to take 15% of notebook market this year
278	0.09	New Intel roadmap reveals Conroe L "solo", mobile plans
300	0.09	Interactive display system knows users by touch
341	0.09	A DNA Database For All U.S. Workers?
540	0.08	Computer Viruses Monitored via Dynamic Worldmap
258	0.08	New Sensor Technology Looks at Molecular 'Fingerprint'
149	0.07	Supreme Court won't consider Yahoo case
247	0.07	Lambda Table - A high-res tiled LCD table and interaction device
642	0.03	Interactive dining table
1204	0.03	Websites as graphs: Visualizing the DOM Structure of Websites
532	0.02	MIT Technology Review Launches New Micro-documentary Video Series