Paper: Using a Model of Social Dynamics to Predict Popularity of News

Authors: Kristina Lerman & Tad Hogg

Presenter: **O**mid **D**avtalab

CSCI 586

Professor McLeod

Spring 2015

Popularity of Content in **Social Media**

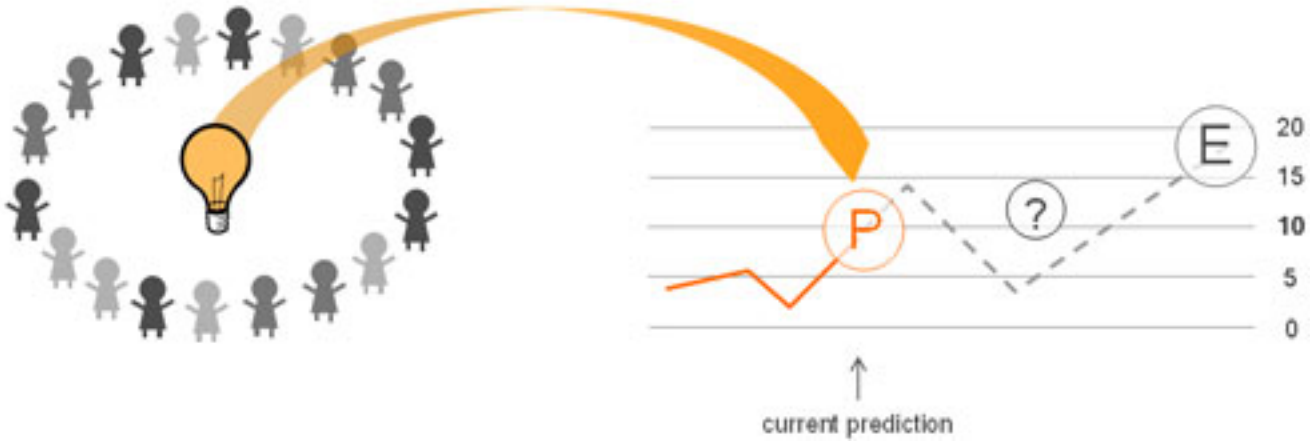
- Popularity of content in social media is unequally distributed, with some items receiving a disproportionate share of attention from users.
- Predicting which newly-submitted items will become popular is critically important for both companies that host social media sites and their users.

Why **Popularity Prediction** is Important?

- ✓ **Accurate** and **timely prediction** would enable the companies to **maximize revenue** through differential pricing for access to **content** or **ad placement**.

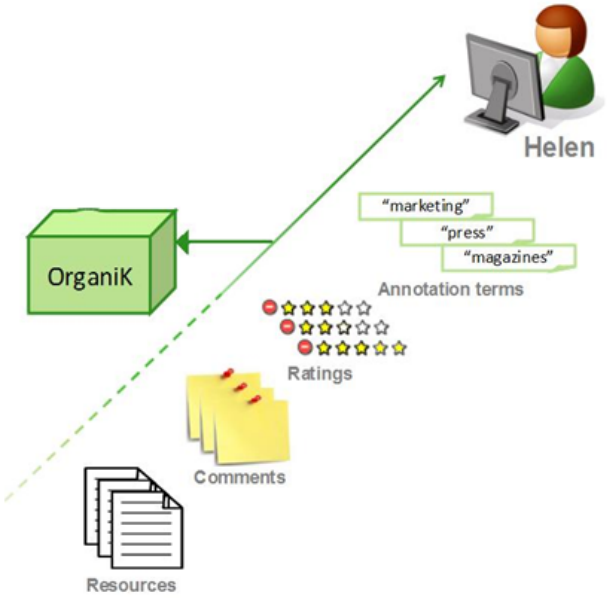
Popularity Prediction

- Predicting popularity of content in social media, however, is challenging due to the complex interactions among **content quality**, how the social media site chooses to **highlight content**, and **influence among users**.



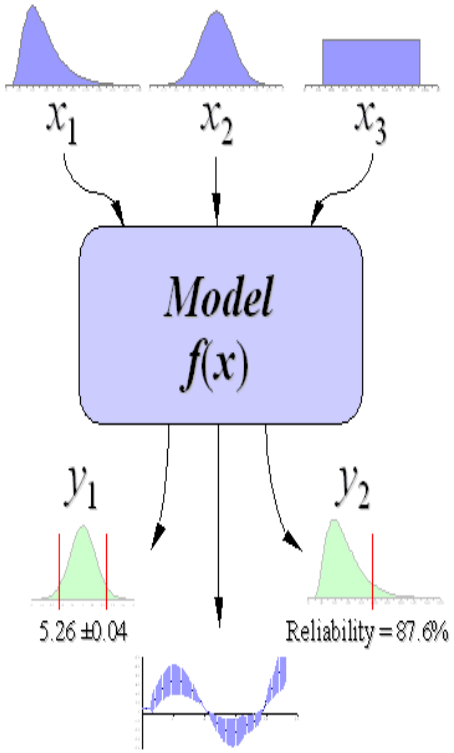
How Can We **Predict**?

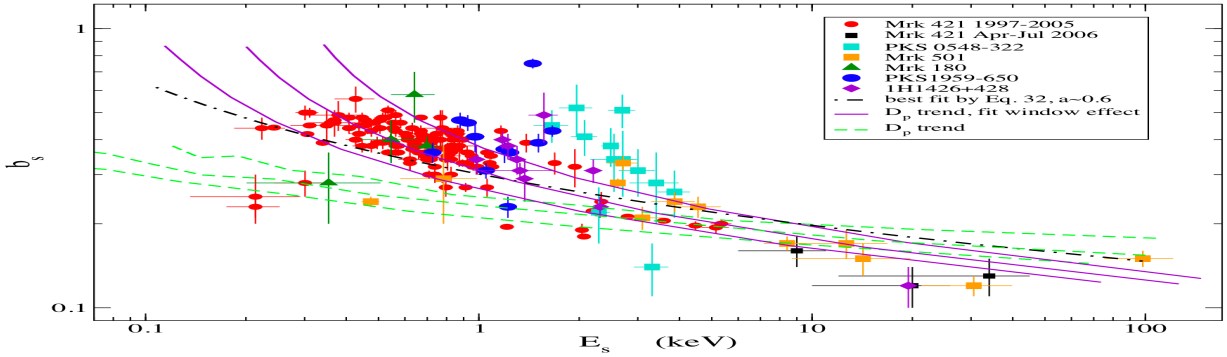
- Using **Stochastic Models** of **user behavior** on these sites allows predicting popularity based on early user reactions to new content.



Stochastic Models for User Behavior Prediction

User response to contributed content in **online social media** depends on **many factors**. These include **how the site lays out new content**, **how frequently the user visits the site**, **how many friends the user follows**, **how active these friends are**, as well as **how interesting or useful the content is** to the user.





Resource for Using Stochastic Models in User Behavior Prediction:
http://arxiv.org/abs/1308.2705

Authors' Claim for Popularity Prediction

Authors claim that **modeling the collective behavior of** users of a **social media** site allows them to **predict the popularity of items** from the **users' early reaction** to them.

They investigate the claim empirically using **data** from the **social news portal** called **Digg**.

http://digg.com

Digg (Social News Portal)

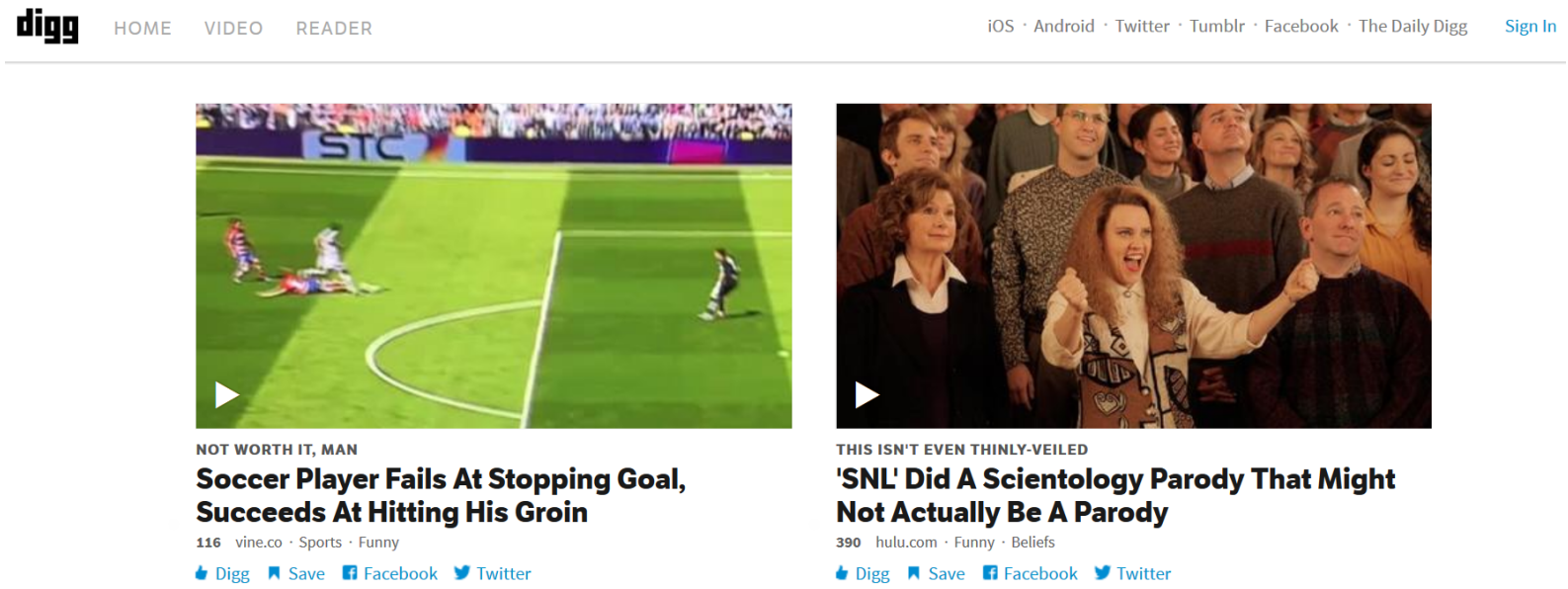
- **Digg** is a news aggregator with an editorially driven front page, aiming to select stories specifically for the Internet audience such as science, trending political issues, and viral Internet issues. It was launched in its current form on July 31, 2012, with support for sharing content to other social platforms such as Twitter and Facebook

Digg, Inc.

Founded	November 2004
Headquarters	New York City, New York, United States[1]
Area served	Worldwide
Founder(s)	Jay Adelson & Kevin Rose[2]
Key people	Matt Williams (CEO)
Revenue	US$8.5 million (2008 est.)[3]
Owner	Betaworks
Employees	11 (2012)[4]
Website	digg.com
Alexa rank	▼ 679 (December 2014)[5]
Type of site	Social news
Advertising	None
Registration	Optional
Available in	English
Launched	December 5, 2004
Current status	Active

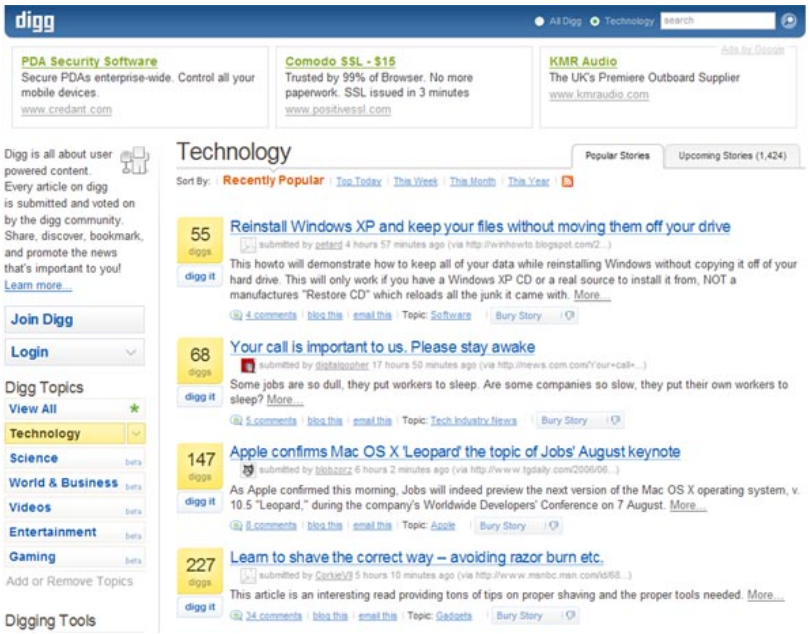
Digg Website

- It formerly had been a very popular social news website, allowing people to vote web content up or down, called *digging* and *burying*, respectively.



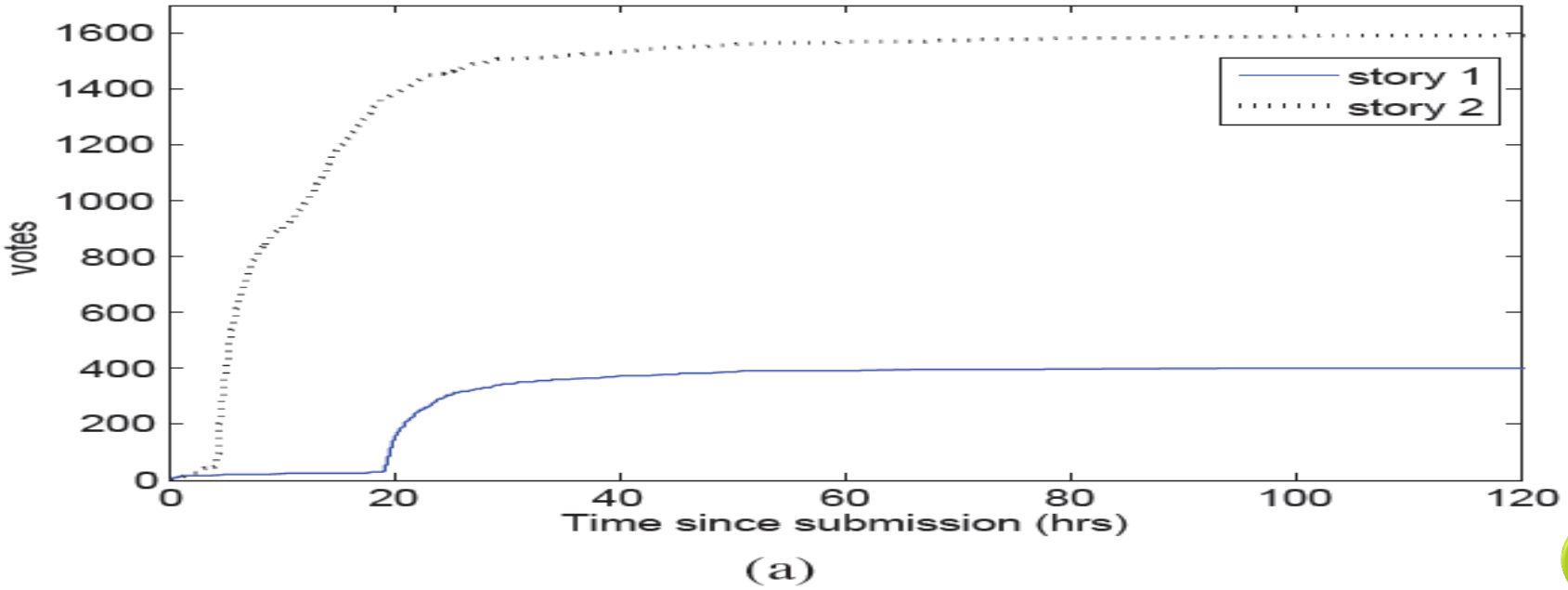
Digg Website

- Digg allows users to submit and rate news stories by voting on, or 'digging', them.
- There are many new submissions every minute, over 16,000 a day.
- Every day Digg picks about a hundred stories that it deems to be ***popular*** and promotes them to the front page.



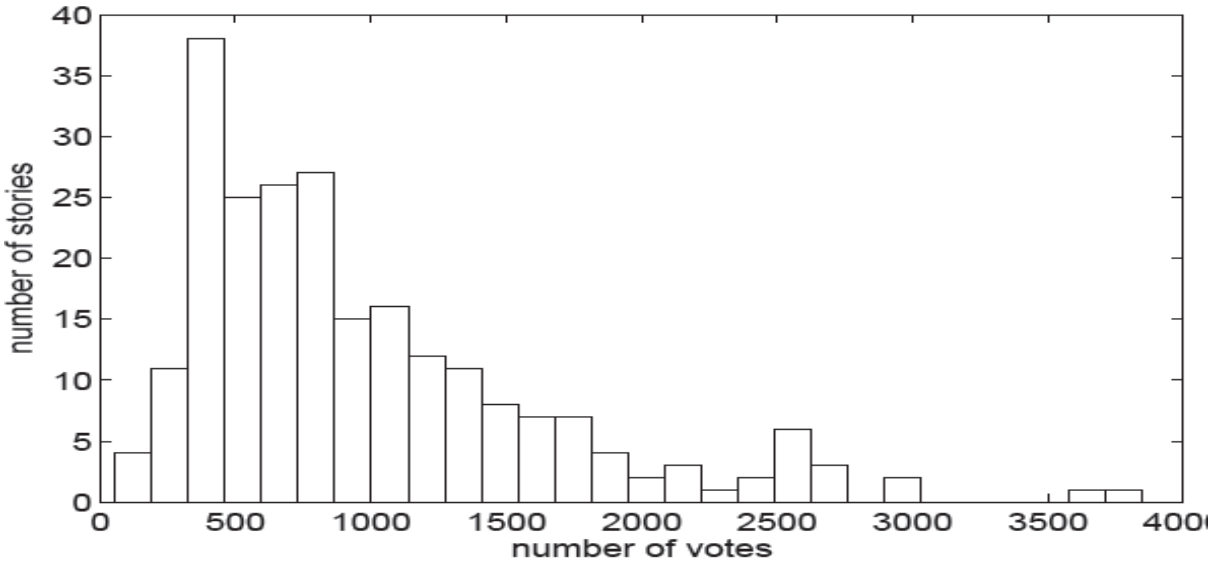
Evolution of the Number of Votes

- While a story is in the upcoming stories list, it **accrues votes slowly**. After it is promoted to the **front page**, it accumulates votes at a **much faster pace**.



InEquality of POpularity

- This Figure shows the **distribution of the final number of votes** received by **front page stories** that were submitted over a period of about two days in June 2006.



(b)

Dynamical Model of Social Voting

- At an aggregate level, we focus on how the number of votes a story receives changes over time.
- The changing state of a story is characterized by three values:

Nvote(t)	the number of votes the story has received by time t after it was submitted to Digg
List	the list the story is in at time t (upcoming or front pages)
Location	its location within that list

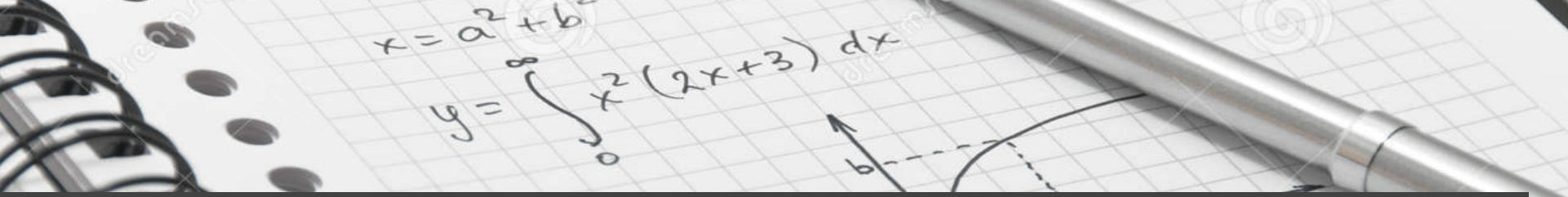
Rate Equation for Nvote(t)

Visibility + Interest

$$\frac{dN_{\text{vote}}(t)}{dt} = r(\nu_{\text{f}}(t) + \nu_{\text{u}}(t) + \nu_{\text{friends}}(t))$$

The votes a story receives depends on the combination of its **visibility** and **interest**, with visibility coming from different parts of the Digg user interface: the , **front and upcoming page lists, friends interface**, and the **position within each list**

r measures how **interesting** the story is, i.e., the **probability a user seeing the story will vote on it**



How to get **visible/promoted** on the upcoming/front stories pages

- ✓ When the number of accumulated votes exceeds a promotion **threshold h**, the story moves to the front page.

$$
\begin{aligned}
\nu_{\mathrm{f}} &= \nu f_{\mathrm{page}}(p(t))\,\Theta(N_{\mathrm{vote}}(t) - h) \\
\nu_{\mathrm{u}} &= c\,\nu f_{\mathrm{page}}(q(t))\,\Theta(h - N_{\mathrm{vote}}(t))\Theta(24\mathrm{hr} - t) \\
\nu_{\mathrm{friends}} &= \omega s(t)
\end{aligned}
$$

- v is the rate users visit Digg.
- The step function (Nvote(t) - h) - when N (t) > h, story is visible on front page
- The step function Θ (24hr - t) , story staying in the upcoming for at most 24hrs

S(t) Model

- ✓ Now, we model s(t), the number of fans of voters on the story by time t who have not yet seen the story:

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

Solved by assuming: $N_{vote}(0) = 1$

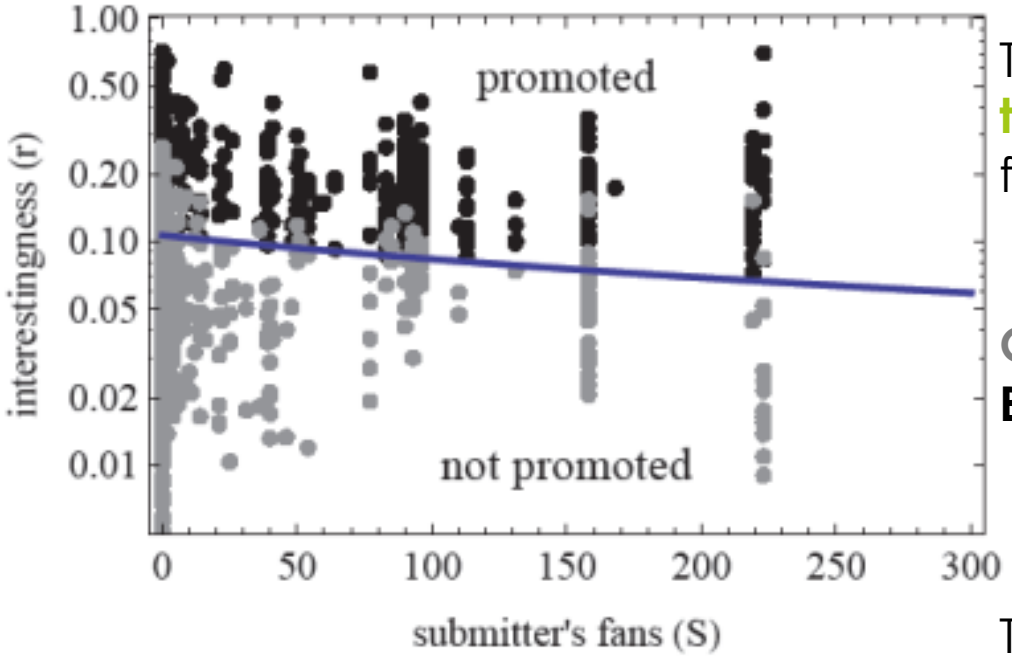
SUBMITTER!

Model Parameters:

parameter	value
rate general users come to Digg	$\nu = 10$ users/min
fraction viewing upcoming pages	$c = 0.3$
rate a voters' fans come to Digg	$\omega = 0.002$/min
page view distribution	$\mu = 0.6, \lambda = 0.6$
fans per new vote	$a = 51, b = 0.62$
vote promotion threshold	$h = 40$
upcoming stories list location	$k_{\mathrm{u}} = 0.06$ pages/min
front page list location	$k_{\mathrm{f}} = 0.003$ pages/min
story specific parameters	
interestingness	r
number of submitter's fans	S

Table 1: Model parameters. Parameters specifying page view distribution are defined in [9].

What does Promotion Function gives us?



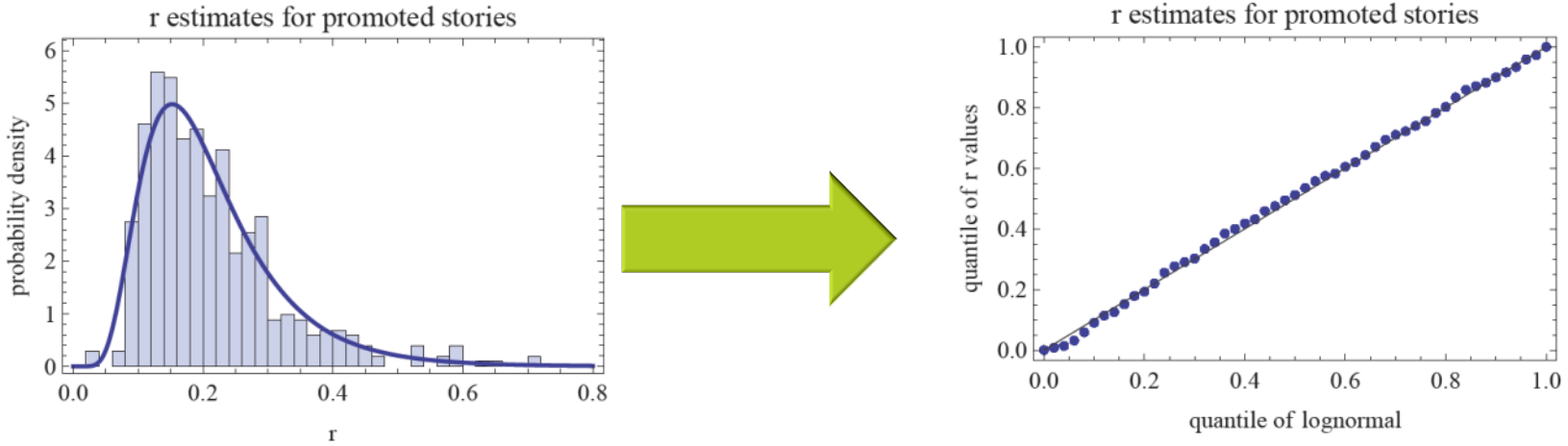
The model predicts stories **above the curve are promoted** to the front page.

Grey -> **Not Promoted**
Black -> **Promoted**

They have **False Positive** points
Accuracy = %95

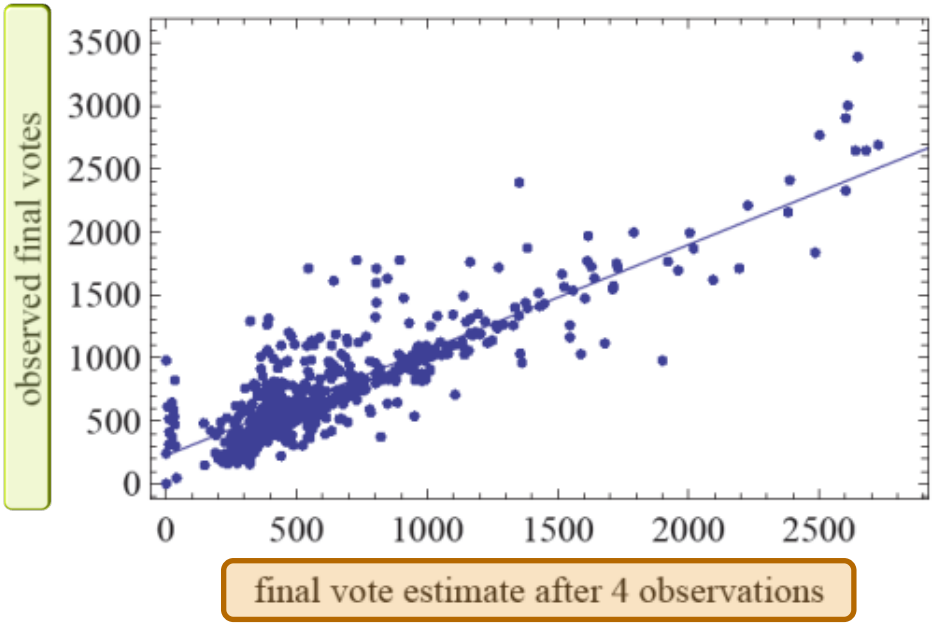
Estimating story quality

- We can estimate how interesting a story is by comparing the model's solutions to the observed popularity of the story.



- Difference between actual number of votes and the predicted number of votes a story receives.

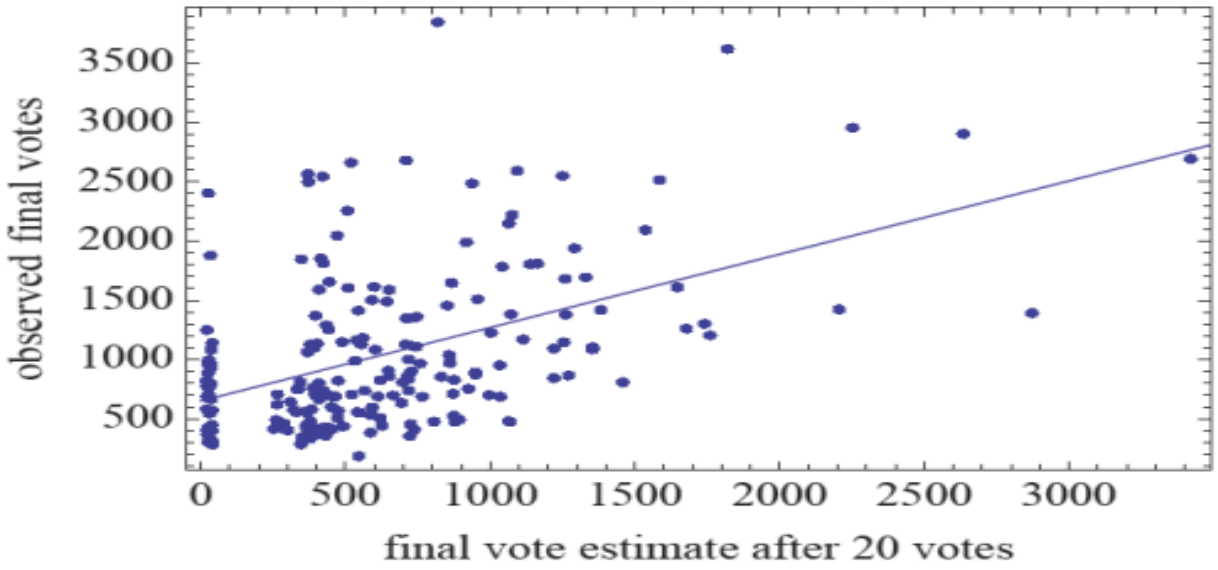
Actual Vs. **Estimated** Final Votes



- Comparing **observed number of final votes** and **estimated number of final votes**.
- The **ideal slope should be 1**, however, here the **slope is 0.84** which comes from the best linear fit

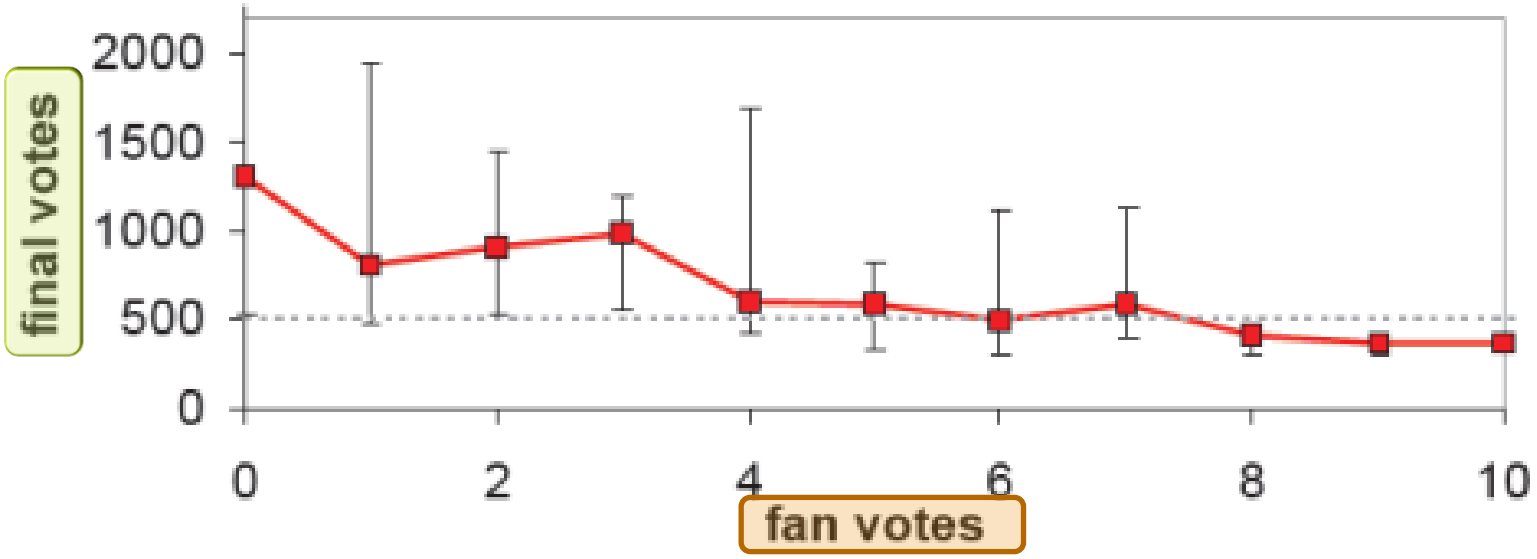
Prediction Model Using **First 10 Votes**

Previous predictions apply to promoted stories only and do not take into account changes in visibility of a story through growth in the number of fans



This simpler model, which does not consider the number of fans for the story's voters, has a lower correlation, 75%.

Last Model: Number of Fans in First 10 Votes



Number of fan votes within the **first 10 votes** vs. **actual final votes** received by front page stories.

Selection of **Stories**

Selection of stories from the May data set with the highest and lowest r values. For each story, we show the final number of votes it received, its estimated r value, and its title.

final votes	estimated r	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

Conclusions

- Stories that receive many fan votes, i.e., votes from fans of the submitter or previous voters, ultimately go on to accumulate fewer votes than stories that initially receive few fan votes.
- Social influence during the early voting period and the final number of votes a story receives are inversely correlated.
- The model makes several assumptions and approximations which could reduce accuracy of prediction (treated promotion as an exact threshold, 40)

Authors' Suggested Future Work

- Social influence offers valuable evidence about story's interest within and outside a community.
- Monitoring the spread of interest in a story through the fan network will lead to a better estimate of r.
- The value of r could be different to fans vs non-fans.

Questions?

Thanks!