Going viral

Matthew J. Salganik

Social Network (Soc 204)
Spring 2017
Princeton University

April 26, 2017

Logistics:

- Final class

Logistics:

- Final class
- Final exam

Vote:

1. Goel, S. et al (2016) "The structural virality of online diffusion." *Management Science*
2. Cheng et al. (2014) "Can cascades be predicted?" *WWW*

What are some examples of things going viral?

- The Harlem Shake
- Ellen Degeneres' Oscar selfie
- Gangam style
- Kim Kardashian's "Break the Internet" magazine cover
- Jay Z elevator fight
- Kanye and taylor swift at the awards
- Various Vines. Often one very popular vine goes viral and is used in many, many (it gets annoying) other vines. (I.e. "Deez Nuts... Gotteem" which is popular right now)
- Just today, there was the quadruple rainbow photo
- Anything involving cute kids, e.g. https://www.youtube.com/watch?v=TP8RB7UZHKI
- Jimmy Fallon's Lip Sync Battles (went viral so frequently that they made an entire TV series out of them)
- #blacklivesmatter
- "Waka Flocka Flame for President" video (today)
- "Wrecking Ball" by Miley Cyrus
- "Hump Day" Geico commercials
- Drake memes

- Both papers deal with a similar empirical phenomena and both struggle to figure out what is the **right question**

- Both papers deal with a similar empirical phenomena and both struggle to figure out what is the **right question**
- Both papers include small and big cascades offering a systematic approach

- Both papers deal with a similar empirical phenomena and both struggle to figure out what is the **right question**
- Both papers include small and big cascades offering a systematic approach
- Both papers are in Pasteur's quadrant (motivated by use and seeking fundamental understanding)

- Both papers deal with a similar empirical phenomena and both struggle to figure out what is the **right question**
- Both papers include small and big cascades offering a systematic approach
- Both papers are in Pasteur's quadrant (motivated by use and seeking fundamental understanding)
- The papers end up with different ways of approaching the problem: descriptive vs predictive

The Structural Virality of Online Diffusion

Sharad Goel, Ashton Anderson
Stanford University, Stanford, California, 94305 {scgoel@stanford.edu, ashton@cs.stanford.edu}
Jake Hofman, Duncan J. Watts
Microsoft Research, New York, New York 10016 {jmh@microsoft.com, duncan@microsoft.com}

What is virality?

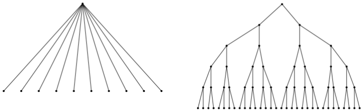


Figure 1 **A schematic depiction of broadcast versus viral diffusion, where nodes represent individual adoptions and edges indicate who adopted from whom.**

Wiener index (from chemistry):

$$\nu(T) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}$$

where $d_{i,j}$ is the length of the shortest path between i and j

In other words, expected path length between two randomly chosen points

What is virality?

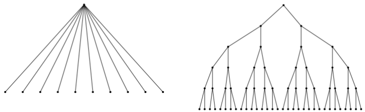


Figure 1 **A schematic depiction of broadcast versus viral diffusion, where nodes represent individual adoptions and edges indicate who adopted from whom.**

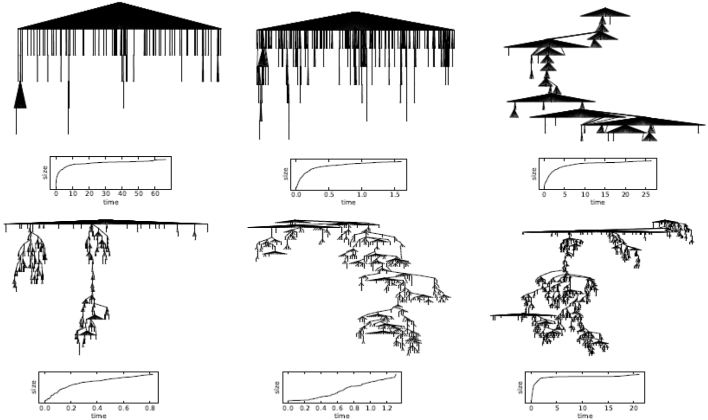


Figure 3 A random sample of cascades stratified and ordered by increasing structural virality, ranging from 2 to 50. For ease of visualization, cascades were restricted to having between 100 and 1000 adopters. Cumulative adoption curves (i.e., total cascade size over time) are shown below each cascade, with time indicated in hours.

describing outcomes vs describing generative process

What do viral cascades look like?

- 622 million unique pieces of content shared via Twitter
- 1.2 billion adoptions
- videos, images, news stories, and petitions

"Big data" is needed because large cascades are very, very rare.

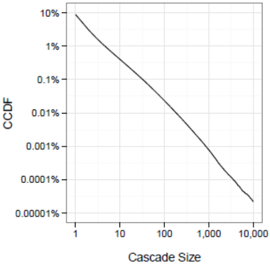


Figure 2 Distribution of cascade sizes on a log-log scale, aggregated across the four domains we study: videos, news, pictures, and petitions.

Most things don't grow, but they focus on the cascades that include at least 100 nodes (1 in 4,000 events).

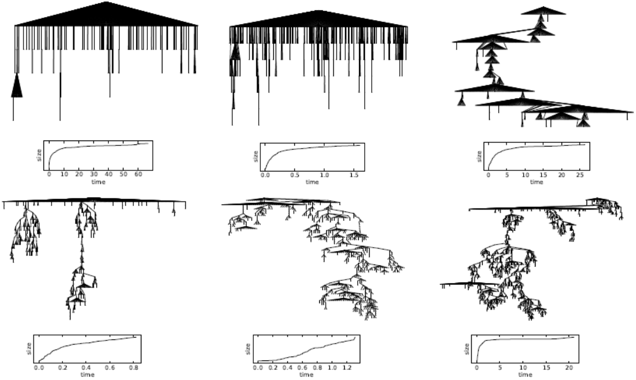


Figure 3 A random sample of cascades stratified and ordered by increasing structural virality, ranging from 2 to 50. For ease of visualization, cascades were restricted to having between 100 and 1000 adopters. Cumulative adoption curves (i.e., total cascade size over time) are shown below each cascade, with time indicated in hours.

Structural virality seems to capture something different from speed of adoption and diffusion curves

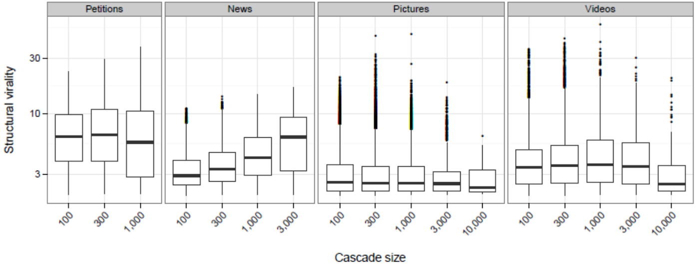


Figure 5 Boxplot of structural virality by size on a log-log scale, separated by domain. Lines inside the boxes indicate median structural virality, while the boxes themselves show interquartile ranges.

knowing the size of a cascades reveals little about its structure

What combination of spreading process and network structure is consistent with these results?

What combination of spreading process and network structure is consistent with these results?
SIR model on network with power law degree distribution

How might the ideas in this paper be used?

Donald J. Trump @realDonaldTrump

Follow

'Presidential Executive Order on Promoting Agriculture and Rural Prosperity in America' Executive Order: 45.wh.gov/UDWud3

RETWEETS	LIKES
3,509	15,493

https://www.youtube.com/watch?v=wSw0szoHuoI

Can cascades be predicted?

Justin Cheng
Stanford University
jcccf@cs.stanford.edu

Lada A. Adamic
Facebook
ladamic@fb.com

P. Alex Dow
Facebook
adow@fb.com

Jon Kleinberg
Cornell University
kleinber@cs.cornell.edu

Jure Leskovec
Stanford University
jure@cs.stanford.edu

Fundamental question:

- given a cascade that current has size k, will grow beyond the median size of $f(k)$?
- given a cascade of size k, will the cascade double in size and reach at least $2k$ nodes?

Two questions (same in this case):

- Is this tweet going to get a lot of retweets?
- Given that this tweet has already had x retweets can I predict if it will get $2x$ retweets?

Takes a machine learning approach (e.g., COS 424)

Takes a machine learning approach (e.g., COS 424)

Content Features	
$score_{food/nature/\ldots}$	The probability of the photo having a specific feature (food, overlaid text, landmark, nature, etc.)
is_en	Whether the photo was posted by an English-speaking user or page
$has\_caption$	Whether the photo was posted with a caption
$liwc_{pos/neg/soc}$	Proportion of words in the caption that expressed positive or negative emotion, or sociality, if English
Root (Original Poster) Features	
$views_{0,\ k}$	Number of users who saw the original photo until the kth reshare was posted
$orig\_is\_page$	Whether the original poster is a page
$outdeg(v_0)$	Friend, subscriber or fan count of the original poster
age_0	Age of the original poster, if a user
$gender_0$	Gender of the original poster, if a user
fb_age_0	Time since the original poster registered on Facebook, if a user
$activity_0$	Average number of days the original poster was active in the past month, if a user
Resharer Features	
$views_{1..k-1,\ k}$	Number of users who saw the first $k-1$ reshares until the kth reshare was posted
$pages_k$	Number of pages responsible for the first k reshares, including the root, or $\sum_{i=0}^{k} 1\{v_i \text{ is a page}\}$
$friends_k^{avg/90p}$	Average or 90th percentile friend count of the first k resharers, or $\frac{1}{k}\sum_{i=1}^{k} outdeg_{friends}(v_i)1\{v_i \text{ is a user}\}$
$fans_k^{avg/90p}$	Average or 90th percentile fan count of the first k resharers, or $\frac{1}{k}\sum_{i=1}^{k} outdeg(v_i)1\{v_i \text{ is a page}\}$
$subscribers_k^{avg/90p}$	Average or 90th percentile subscriber count of the first k resharers, or $\frac{1}{k}\sum_{i=1}^{k} outdeg_{subscriber}(v_i)1\{v_i \text{ is a user}\}$
$fb\_ages_k^{avg/90p}$	Average or 90th percentile time since the first k resharers registered on Facebook, or $\frac{1}{k}\sum_{i=1}^{k} fb\_age_i$
$activities_k^{avg/90p}$	Average number of days the first k resharers were active in July, or $\frac{1}{k}\sum_{i=1}^{k} activity_i$
$ages_k^{avg/90p}$	Average age of the first k resharers, or $\frac{1}{k}\sum_{i=1}^{k} age_i$
$female_k$	Number of female users among the first k resharers, or $\sum_{i=1}^{k} 1\{gender_i \text{ is female}\}$
Structural Features	
$outdeg(v_i)$	Connection count (sum of friend, subscriber and fan counts) of the ith resharer (or out-degree of v_i on $G=(V,E)$)
$outdeg(v_i')$	Out-degree of the ith reshare on the induced subgraph $G'=(V',E')$ of the first k resharers and the root
$outdeg(\hat{v}_i)$	Out-degree of the ith reshare on the reshare graph $\hat{G}=(\hat{V},\hat{E})$ of the first k reshares
$orig\_connections_k$	Number of first k resharers who are friends with, or fans of the root, or $\lvert\{v_i \mid (v_0,v_i)\in E,\ 1\le i\le k\}\rvert$
$border\_nodes_k$	Total number of users or pages reachable from the first k resharers and the root, or $\lvert\{v_i \mid (v_i,v_j)\in E,\ 0\le i,j\le k\}\rvert$
$border\_edges_k$	Total number of first-degree connections of the first k resharers and the root, or $\lvert\{(v_i,v_j) \mid (v_i,v_j)\in E,\ 0\le i,j\le k\}\rvert$
$subgraph_k'$	Number of edges on the induced subgraph of the first k resharers and the root, or $\lvert\{(v_i,v_j) \mid (v_i,v_j)\in E',\ 0\le i,j\le k\}\rvert$
$depth_k'$	Change in tree depth of the first k reshares, or $\min_\beta \sum_{i=1}^{k}(depth_i - \beta i)^2$
$depths_k^{avg/90p}$	Average or 90th percentile tree depth of the first k reshares, or $\frac{1}{k}\sum_{i=1}^{k} depth_i$
did_leave	Whether any of the first k reshares are not first-degree connections of the root
Temporal Features	
$time_i$	Time elapsed between the original post and the ith reshare
$time_{1..k/2}'$	Average time between reshares, for the first $k/2$ reshares, or $\frac{1}{k/2-1}\sum_{i=1}^{k/2-1}(time_{i+1}-time_i)$
$time_{k/2..k}'$	Average time between reshares, for the last $k/2$ reshares, or $\frac{1}{k/2-1}\sum_{i=k/2}^{k-1}(time_{i+1}-time_i)$
$time_{1..k}''$	Change in the time between reshares of the first k reshares, or $\min_\beta \sum_{i=1}^{k-1}(time_{i+1}-time_i)-\beta i)^2$
$views_{0,k}'$	Number of users who saw the original photo, until the kth reshare was posted, per unit time, or $\frac{views_{0,\ k}}{time_k}$
$views_{1..k-1,\ k}'$	Number of users who saw the first $k-1$ reshares, until the kth reshare was posted, per unit time, or $\frac{views_{1..k-1,k}}{time_k}$

Table 1: List of features used for learning. We compute these features given the cascade until the kth reshare.

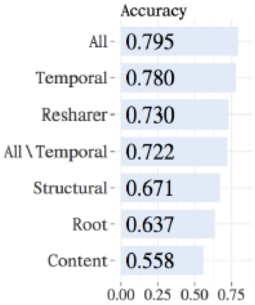


Figure 4: Using logistic regression, we are able to predict with near 80% accuracy whether the size of a cascade will reach the median (10) after observing the first k =5 reshares.

Temporal features are most important

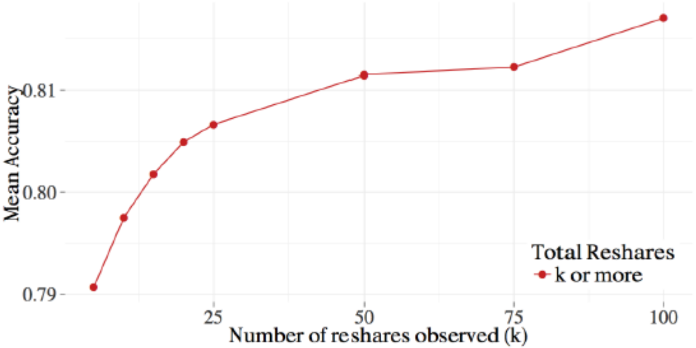
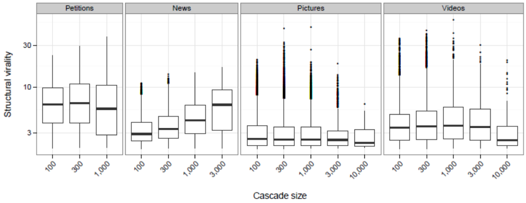
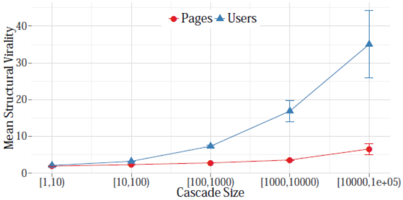


Figure 5: If we observe the first k reshares of a cascade, and want to predict whether the cascade will double in size, our prediction improves as we observe more of it.

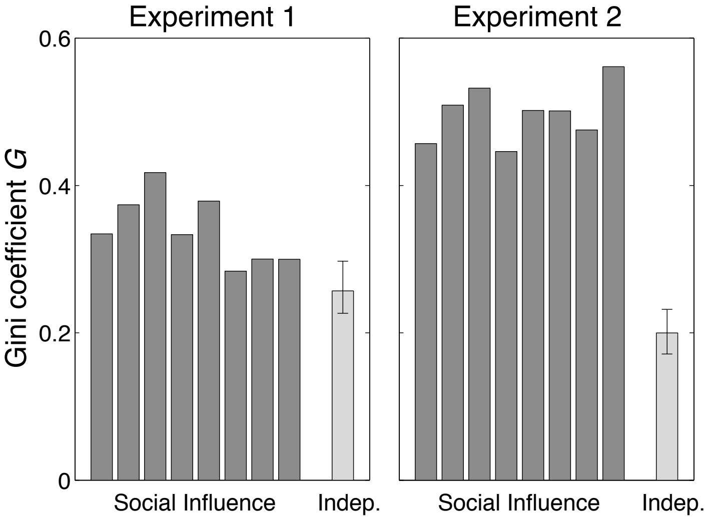
Cascades become slightly more predictable over time


Petitions
News
Pictures
Videos
Structural virality
30
10
3
100
300
1,000
3,000
10,000
Cascade size


Pages
Users
Mean Structural Virality
40
30
20
10
0
[1,10)
[10,100)
[100,1000)
[1000,10000)
[10000,1e+05)
Cascade Size


SUP BRO

gini coefficient: 0.787!


Experiment 1
Experiment 2
Gini coefficient G
0.6
0.4
0.2
0
Social Influence
Indep.
Social Influence
Indep.

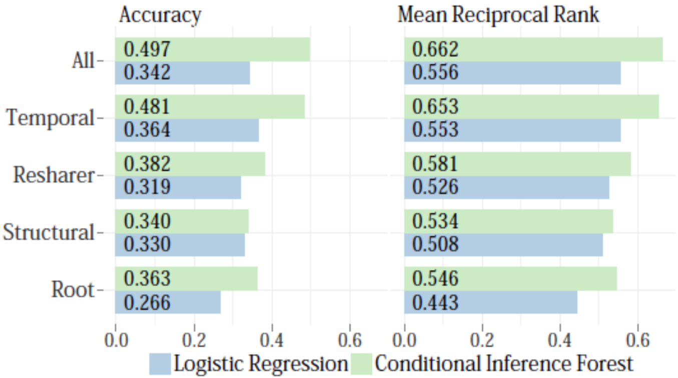


Figure 10: In predicting the largest cascade in clusters of 10 or more cascades of identical photos, we perform significantly above the baseline of 0.1.

What are some examples of things going viral?

- The Harlem Shake
- Ellen Degeneres' Oscar selfie
- Gangam style
- Kim Kardashian's "Break the Internet" magazine cover
- Jay Z elevator fight
- Kanye and taylor swift at the awards
- Various Vines. Often one very popular vine goes viral and is used in many, many (it gets annoying) other vines. (I.e. "Deez Nuts... Gotteem" which is popular right now)
- Just today, there was the quadruple rainbow photo
- Anything involving cute kids, e.g. https://www.youtube.com/watch?v=TP8RB7UZHKI
- Jimmy Fallon's Lip Sync Battles (went viral so frequently that they made an entire TV series out of them)
- #blacklivesmatter
- "Waka Flocka Flame for President" video (today)
- "Wrecking Ball" by Miley Cyrus
- "Hump Day" Geico commercials
- Drake memes

Summary:

- almost nothing posted on Twitter and Facebook creates a large cascades

Summary:

- almost nothing posted on Twitter and Facebook creates a large cascades
- tweets and photos from FB pages show little relationship between structural virality and cascades size; photos from FB users that create large cascades are structurally viral

Summary:

- almost nothing posted on Twitter and Facebook creates a large cascades
- tweets and photos from FB pages show little relationship between structural virality and cascades size; photos from FB users that create large cascades are structurally viral
- there are many different ways to ask interesting questions about going viral

http://bit.ly/socnet204

http://bit.ly/socnet204

Face to face networks and the spread of disease

Some diseases spread through face-to-face contact. How can we measure face-to-face contact networks?

- **surveys**
- "sociotechnical networks" (e.g., Twitter, Facebook, email, etc)
- mobile phones (e.g., Bluetooth scans such as Eagle et al)
- **wearable sensors**