Class 17: Going viral

Matthew J. Salganik

Sociology 204: Social Networks
Princeton University

2/2 Can cascades be predicted?

Can cascades be predicted?

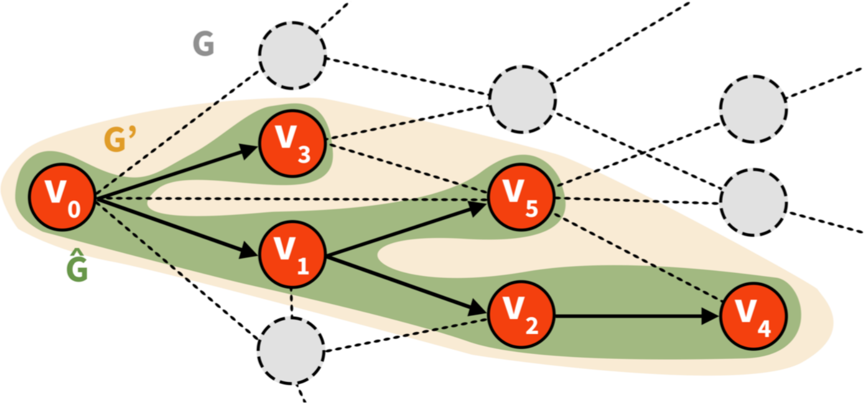
Justin Cheng
Stanford University
jcccf@cs.stanford.edu

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



Reshare cascades of images on Facebook in June 2013

Two ways of posting the same question (in this case):

- 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?

We therefore propose the following *cascade growth prediction problem*: given a cascade that currently has size k, predict whether it grow beyond the median size $f(k)$. (As we show later, the prediction problem is equivalent to asking: given a cascade of size k, will the cascade double its size and reach at least $2k$ nodes?) This implicitly defines a family of prediction problems, one for each k. We can thus ask how cascade predictability behaves as we sweep over larger and larger values of k. (There are natural variants and generalizations in which we ask about reaching target sizes other than the median $f(k)$.) This problem formulation has a number of strong advantages over standard ways of trying to define cascade prediction. First, it leads to a prediction problem in which the classes are balanced, rather than highly unbalanced. Second, it allows us to ask for the first time how the predictability of a cascade varies over the range of its growth from small to large. Finally, it more closely approximates the real tasks that need to be solved in applications for managing viral content, where many evolving cascades are being monitored, and the question is which are likely to grow significantly as time moves forward.

We therefore propose the following *cascade growth prediction problem*: given a cascade that currently has size k, predict whether it grow beyond the median size $f(k)$. (As we show later, the prediction problem is equivalent to asking: given a cascade of size k, will the cascade double its size and reach at least $2k$ nodes?) This implicitly defines a family of prediction problems, one for each k. We can thus ask how cascade predictability behaves as we sweep over larger and larger values of k. (There are natural variants and generalizations in which we ask about reaching target sizes other than the median $f(k)$.) This problem formulation has a number of strong advantages over standard ways of trying to define cascade prediction. First, it leads to a prediction problem in which the classes are balanced, rather than highly unbalanced. Second, it allows us to ask for the first time how the predictability of a cascade varies over the range of its growth from small to large. Finally, it more closely approximates the real tasks that need to be solved in applications for managing viral content, where many evolving cascades are being monitored, and the question is which are likely to grow significantly as time moves forward.

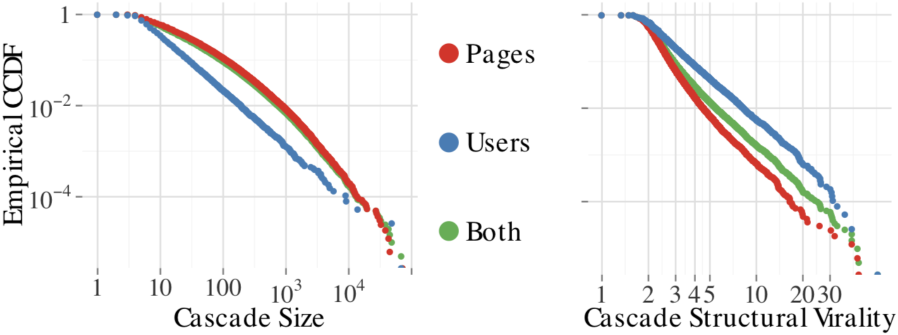
Why might we need to manage viral content?

- Amplify virality

We therefore propose the following *cascade growth prediction problem*: given a cascade that currently has size k, predict whether it grow beyond the median size $f(k)$. (As we show later, the prediction problem is equivalent to asking: given a cascade of size k, will the cascade double its size and reach at least $2k$ nodes?) This implicitly defines a family of prediction problems, one for each k. We can thus ask how cascade predictability behaves as we sweep over larger and larger values of k. (There are natural variants and generalizations in which we ask about reaching target sizes other than the median $f(k)$.) This problem formulation has a number of strong advantages over standard ways of trying to define cascade prediction. First, it leads to a prediction problem in which the classes are balanced, rather than highly unbalanced. Second, it allows us to ask for the first time how the predictability of a cascade varies over the range of its growth from small to large. Finally, it more closely approximates the real tasks that need to be solved in applications for managing viral content, where many evolving cascades are being monitored, and the question is which are likely to grow significantly as time moves forward.

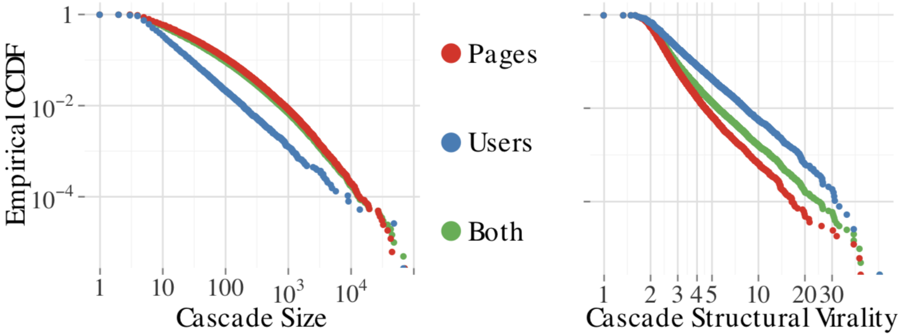
Why might we need to manage viral content?

- Amplify virality
- Check and possible pull things that appear likely to go viral



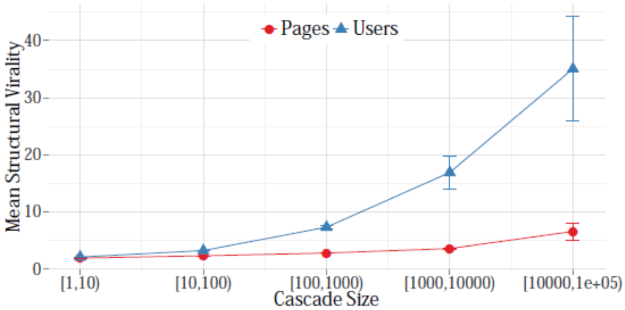
Difference between pages (media, celebrities) and users (organic):

- user cascades are small than page cascades

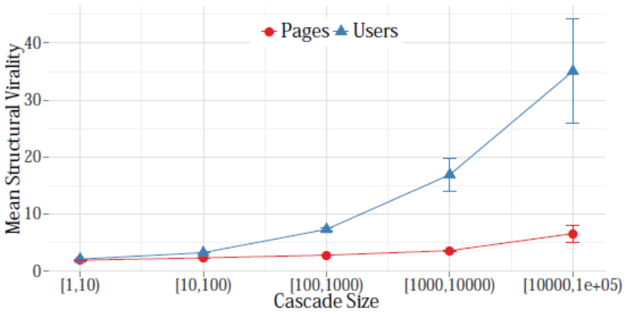


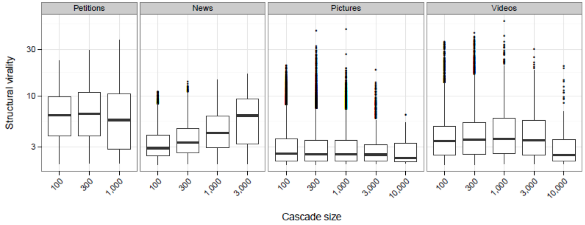
Difference between pages (media, celebrities) and users (organic):

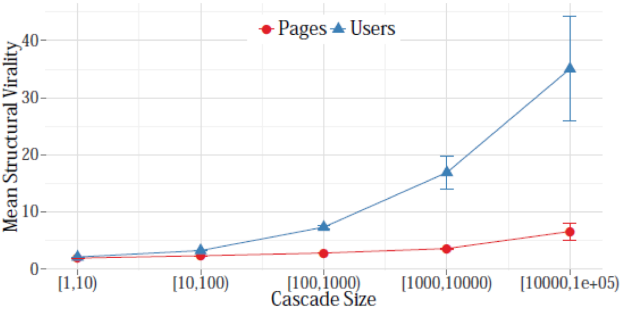
- user cascades are small than page cascades
- user cascades tend to have higher structural virality



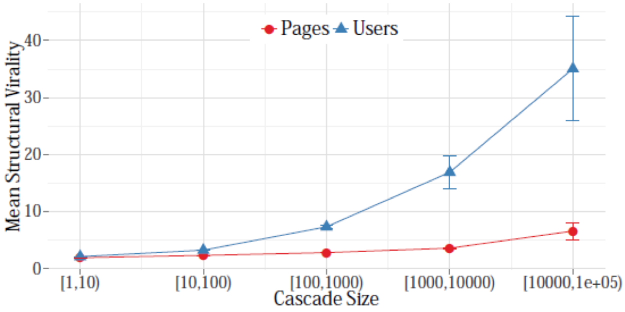
Pages
Users
Mean Structural Virality
0
10
20
30
40
[1,10)
[10,100)
[100,1000)
[1000,10000)
[10000,1e+05)
Cascade Size


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


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



- For page cascades, there is a weak relationship between size and virallity (similar to Goel et al)



- For page cascades, there is a weak relationship between size and virallity (similar to Goel et al)
- For user cascades, there is a strong positive relationship between size and virallity (different to Goel et al)

Machine learning approach (e.g., COS 424) to predicting if a cascade will double

Machine learning approach (e.g., COS 424) to predicting if a cascade will double

Content Features	
$score_{food/nature/...}$	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 $\sum_{i=0}^{k} \mathbb{1}\{v_i \text{ is a page}\}$
$friends_k^{avg/90p}$	Average or 90th percentile friend count of the first k resharers, or $\frac{1}{k}\sum_{i=1}^{k} outdeg_{friends}(v_i)\mathbb{1}\{v_i \text{ is a user}\}$
$fans_k^{avg/90p}$	Average or 90th percentile fan count of the first k resharers, or $\frac{1}{k}\sum_{i=1}^{k} outdeg(v_i)\mathbb{1}\{v_i \text{ is a page}\}$
$subscribers_k^{avg/90p}$	Average or 90th percentile subscriber count of the first k resharers, or $\frac{1}{k}\sum_{i=1}^{k} outdeg_{subscriber}(v_i)\mathbb{1}\{v_i \text{ is a user}\}$
$fb\_ages_k^{avg/90p}$	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 $\sum_{i=1}^{k} \mathbb{1}\{gender_i \text{ is female}\}$
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 $\frac{views_{1..k-1,k}}{time_k}$

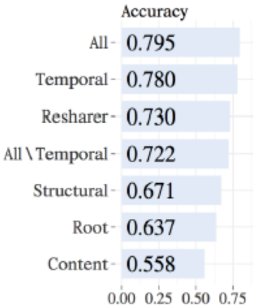


Figure 4: Using logistic regression, we are able to predict with near 80% accuracy whether the size of a cascade will reach the median (10) after observing the first k =5 reshares.

- Temporal features are most predictive (things that spreading fast are likely to keep spreading)

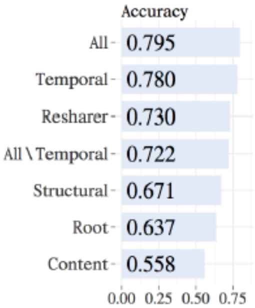


Figure 4: Using logistic regression, we are able to predict with near 80% accuracy whether the size of a cascade will reach the median (10) after observing the first k =5 reshares.

- Temporal features are most predictive (things that spreading fast are likely to keep spreading)
- Content features least predictive

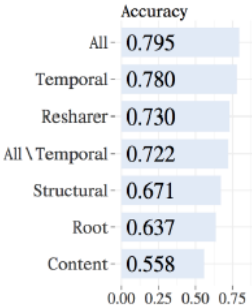


Figure 4: Using logistic regression, we are able to predict with near 80% accuracy whether the size of a cascade will reach the median (10) after observing the first k =5 reshares.

- Temporal features are most predictive (things that spreading fast are likely to keep spreading)
- Content features least predictive
- Temporal features are more predictive than everything else put together

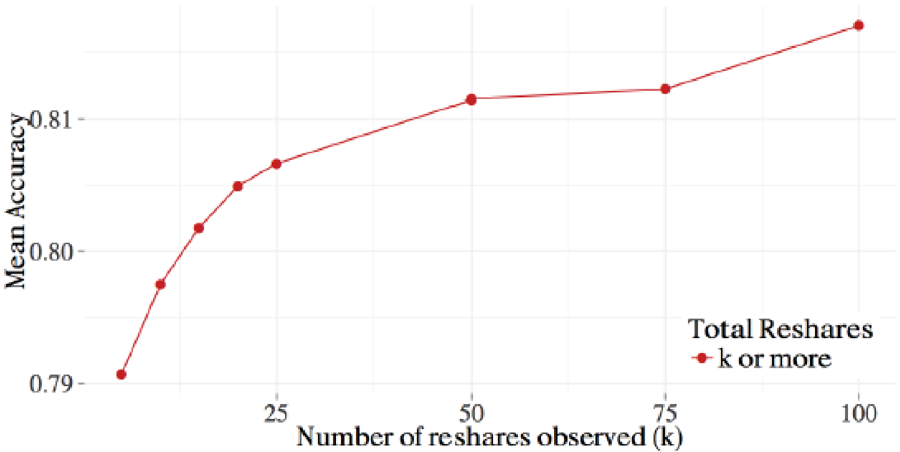
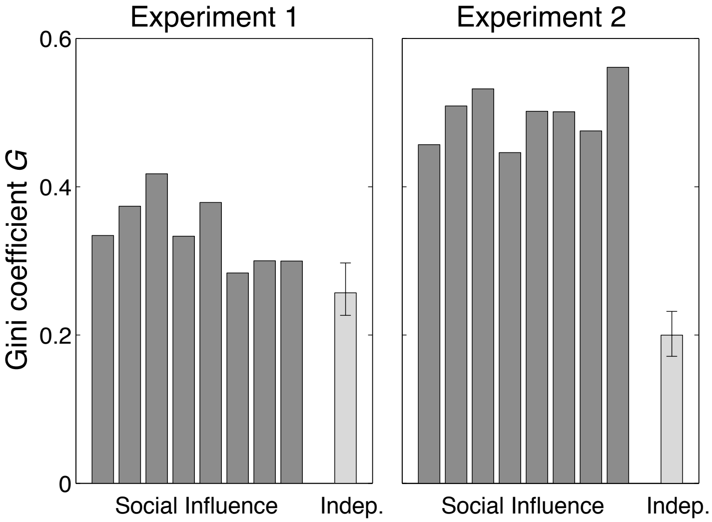


Figure 5: If we observe the first k reshares of a cascade, and want to predict whether the cascade will double in size, our prediction improves as we observe more of it.


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.

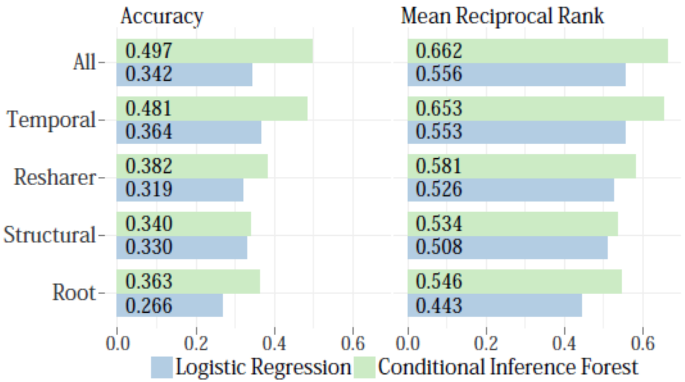


Figure 10: In predicting the largest cascade in clusters of 10 or more cascades of identical photos, we perform significantly above the baseline of 0.1.

We can somewhat predict which of the identical seeds will spread, if we observe the beginning of each cascade

Summary:

- asking the right question can be very important in research

Summary:

- asking the right question can be very important in research
- almost nothing posted on Twitter and Facebook creates a large cascades

Summary:

- asking the right question can be very important in research
- 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:

- asking the right question can be very important in research
- 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

What is all this stuff going viral?

- Kross, E. et al. (2020). Social media and well-being: Pitfalls, progress, and next steps. *Trends in Cognitive Science*.
- Carey, B. (2019). This is your brain off Facebook. *New York Times*.
- Allcott, H. et al. (2020). The welfare effects of social media. *American Economic Review*.
- Baym, N.K. et al. (2020). Mindfully scrolling: Rethinking Facebook after time deactived. *Social Media + Society*.