Class 17: Going viral

Matthew J. Salganik

Sociology 204: Social Networks Princeton University

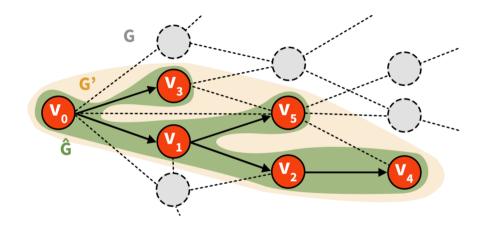
2/2 Can cascades be predicted?



Can cascades be predicted?

Justin Cheng Stanford University jcccf@cs.stanford.edu Lada A. Adamic Facebook Iadamic@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):

nodes?

▶ Given a cascade that current has size k, will grow beyond the median size of f(k)?

 \triangleright Given a cascade of size k, will the cascade double in size and reach at least 2k

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.

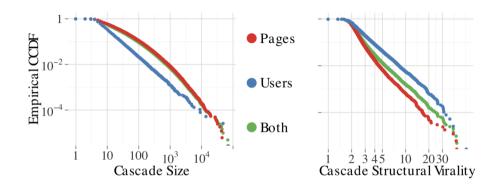
Why might we need to manage viral content?

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.

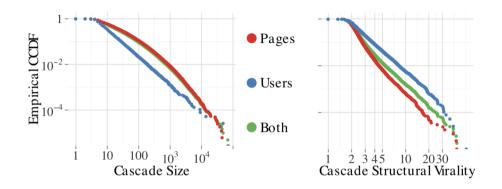
Why might we need to manage viral content?

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



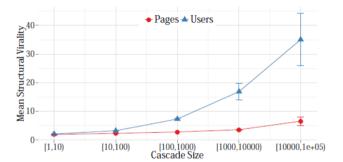
Difference between pages (media, celebrities) and users (organic):

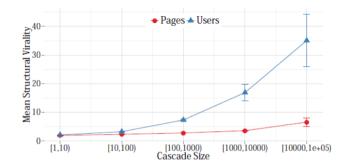
user cascades are small than page cascades

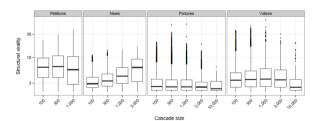


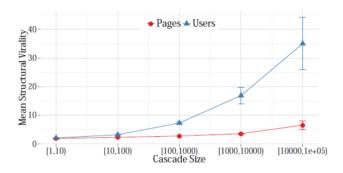
Difference between pages (media, celebrities) and users (organic):

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

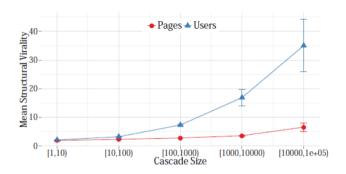








► For page cascades, there is a weak relationship between size and virallity (similar to Goel et al)



- ► 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

Inter-generation Whether the photo was posted with a caption in the caption in the caption of woods in the caption that expressed positive or negative emotion, or sociality, if English interests, a bamber of users who was negating those until the Aft reshare was posted oring, 1s, 1s, 1s, 1s oring, 1s, 1s, 1s, 1s Whether the original poster is a page outledge(us) oring, 1s, 1s, 1s Friend, wherefire or fin count of the original poster or generating one open of the original poster, if a user Gender of the original poster, if a user activity Average number of dues the original poster was active in the past month, if a user views, 1s, 1s Number of users usersomble for the first Postures Number of users usersomble for the first versus, including the root, or S ^{2st} , 1fro is a nave?	
Mether the photos was posted by an flaglish-speaking user or page Whether the photos was posted with a caption Inter_put/may/met Whether the photos was posted with a caption Inter_put/may/met Proportion of words in the caption that expressed positive or negative emotion, or sociality, if English Nest Original Poster Features Nouther of users who sure the eriginal photos until the 4th reshare was posted Whether the original poster is a page Original Poster Features Whether the original poster is a page Priend, subscriber or fan count of the original poster geneters Original Poster Posters Original Poster Posters Resharer Features Number of days the original poster was active in the past month, if a user Resharer Features Number of users who saw the first 6 - 1 reshares suntil the 4th reshare was posted Number of users resonable for the first of reshame, including the responsible to the control or the past month, if a user Number of users resonable for the first or through, include the control or the control	
has _caption Whether the photo was posted with a caption Inter_pust trapt_trapt_ Inter_pust trapt_trapt_ Inter_pust trapt_trapt_ Inter_pust trapt_trapt_ Inter_pust trapt_trapt_ Number of users who saw the cripting those until the Ath reduze was posted Whether the original poster is a page outledge(us) Friend, whether the cripting poster rate page age of the cripting poster, if a user gentlerg Dengte Time since the original poster rate paster of Paccheost, if a user activating Average number of days the cripting poster rate paster of Paccheost, if a user Whether the cripting poster rate paster of Paccheost, if a user Whether the cripting poster rate paster of Paccheost, if a user Whether the cripting poster rate paster of Paccheost, if a user Whether the cripting poster rate paster of Paccheost, if a user	
Proportion of words in the caption that expressed positive or regainse emotion, or sociality, if English Social Clerks, 1	
Root (Original Poster) Features Number of suces who saw the eigening photo until the 4th reduze was posted oring_st_page. Whether the original poster is a page outledge(u) Friend, subscriber or fin count of the original poster genetic genetic Gender of the original poster, if a suer Gender of the original poster, if a suer Them such the original poster registered on Facebook, if a user activating Average number of days the original poster was active in the past menth, if a user Resture Features Number of users secondle for the first Features until the 6th reduce was posted where of the past composite for the first explane, including the root, or \$\frac{5}{2}\ldots\ldo	
streams, b. Number of users who saw the criginal photo until the Ath reduce was posted order, st., page outdog (vs). Whether the criginal poster is a page outdog (vs). Priend, subscriber or fan count of the original poster gender gender Gender of the original poster, if a user Die activation Average number of days the original poster registered or Buchetosk, if a user Average number of days the original poster was active in the past month, if a user Resture Features Number of users secondible of the rich seches sound the daft reduce was posted number of users secondible for the risk reduces, such as most on CVS If (s is a more)	
orig_d_n_pope Whether the original poster is a page outdoug(vit) Pitend, subscriber or in count of the original poster april Age of the original poster, if a user Gender of the original poster, if a user Poster of the original poster, if a user The success to engine poster, if a user Activating Activating Average number of days the original poster was active in the past month, if a user The success to engine poster restores The success to the past of the past month, if a user The success to the past of the past month, if a user The success the past of the past month, if a user Number of the past month for the past month is the past month. If to it is a mare) Whether of the past month for the past month is the past month. If to it is a mare)	
outdog (vs) Friend, subscriber or fan count of the original poure gender g g gender g g gender g g g g g g g g g g g g g g g g g g g	
apen Age of the original poster, if a user gender of the riginal poster, if a user β _− apen Time since the riginal poster is a user in the past month, if a user activity Average number of days the original poster was active in the past month, if a user victory 1 _− 1 _− 1 _− 1 _− 1 _− Number of the second by the original poster was active in the past month, if a user victory 1 _− 1 _− 1 _− 1 _− 1 _− Number of the second by the original poster is a user in the past month, if a user is the past month is a user in the past month, if a user is the past month is a user in the past month in the past month, if a user is the past month is a user in the past month in the past month, if a user is the past month is a user in the past month in the past month, if a user is the past month is a user in the past month in the past month, if a user is the past month is a user in the past month in the past month, if a user is the past month is a user in the past month in the past month, if a user is the past month in the past month in the past month, if a user is the past month in the past mo	
gender 0 Gender of the original power, if a sure D_n ngng	
β _i , aprop. Time since the original poster registered on Facebrook, if a user	
activity_0 Average number of days the original poster was active in the past month, if a user Resharer Features Victors $_{1k-1k}$ Number of users who saw the first $k-1$ reshares until the k th reshare was posted possess. Number of passes responsible for the first reshares, including the root, or Σ^{N}_{k} I (v_i is a page)	
Views _{1k-1.k} Number of users who saw the first $k-1$ reshares until the k th reshare was posted nones. Number of pages responsible for the first k -the first k -the same continues to k -the k -th k	
$views_{1k-1, k}$ Number of users who saw the first $k-1$ reshares until the k th reshare was posted pages. Number of pages responsible for the first k reshares, including the root, or $\sum_{i=0}^{k} 1\{v_i \text{ is a page}\}$	
$pages_k$ Number of pages responsible for the first k reshares, including the root, or $\sum_{i=0}^{k} 1\{v_i \text{ is a page}\}\$ $friends_k^{avy/90p}$ 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 page}\}\$	
$\begin{array}{ll} pages_k & \text{Number of pages responsible for the first k reshares, including the root, or $\sum_{i=0}^k 1 \{v_i$ is a page}\} \\ friends_k^{any/00p} & \text{Average or 90th percentile friend count of the first k resharers, or $\sum_{i=0}^k 1 \{v_i$ is a page}\} \\ & \text{Average or 90th percentile friend count of the first k resharers, or $\sum_{i=0}^k 1 \{v_i$ is a page}\} \\ & \text{Average or 90th percentile friend count of the first k resharers, or $\sum_{i=0}^k 1 \{v_i$ is a page}\}. \end{array}$	
friends k Average or 90th percentile friend count of the first k resharers, or $\frac{1}{k}\sum_{i=1}^{k} outdeg_{friends}(v_i)1\{v_i \mid v_i \mid$	
g = any / 80 p	s a user}
jans _k Average or 9000 percentile ran count of the first k resharers, or	
)1{v _i is a user}
$fb_ages_k^{avg/90p}$ Average or 90th percentile time since the first k resharers registered on Facebook, or $\frac{1}{k}\sum_{i=1}^{k} fb_ag$	e,
$activities_k^{ovg/90p}$ Average number of days the first k resharers were active in July, or $\frac{1}{k}\sum_{i=1}^{k} activity_i$	
$ages_k^{aug/90p}$ 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 <i>i</i> th resharer (or out-degree of v_i or	m G = (V, E)
$outdeg(v'_i)$ Out-degree of the <i>i</i> th reshare on the induced subgraph $G' = (V', E')$ of the first k resharers and the	root
$outdeg(\hat{v}_i)$ Out-degree of the <i>i</i> th 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 $ \{v_i \mid (v_0, v_i) \in E, 1 \le i \le i \le i \le i \le k\}$	{ k}
$border_nodes_k$ Total number of users or pages reachable from the first k resharers and the root, or $ \{v_i \mid (v_i, v_j) \in a_i\} $	$E, 0 \le i, j \le k\}$
$border_edges_k$ Total number of first-degree connections of the first k resharers and the root, or $ \{(v_i, v_j) \mid (v_i, v_j)\} $	$\in E, 0 \le i, j \le k\}$
$subgraph'_k$ Number of edges on the induced subgraph of the first k resharers and the root, or $ \{(v_i, v_j) \mid (v_i, v$	$) \in E', 0 \le i, j \le k$
$depth'_k$ Change in tree depth of the first k reshares, or $\min_{\beta} \sum_{i=1}^{k} (depth_i - \beta i)^2$	
$depths_k^{avy/90p}$ 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; Time elapsed between the original post and the ith reshare	
$\lim_{1,k/2}$ Average time hetween reshares, for the first $k/2$ reshares, or $\frac{1}{k/2-1}\sum_{i=1}^{k/2-1}\{(ime_{i+1}-time_i)\}$ $\lim_{k/2,k}$ Average time hetween reshares, for the last $k/2$ reshares or $\frac{1}{k/2-1}\sum_{i=1}^{k/2-1}\{(ime_{i+1}-time_i)\}$ $\lim_{k/2-1}\sum_{i=1}^{k/2-1}\{(ime_{i+1}-time_i)\}$. Change in the time between reshares of the first k reshares, or $\min_{k/2}\sum_{i=1}^{k/2}(ime_{i+1}-time_i)-[ime_{i+1}-time_i)$.	
$time_{k/2k}^{i,n/s}$ Average time between reshares, for the last $k/2$ reshares, or $\frac{k/2-1}{k/2-1}\sum_{i=k/2}^{k-1}(time_{i+1}-time_i)$	
$time_{1k}^{R/2R}$ Change in the time between reshares of the first k reshares, or $\min_{\beta} \sum_{i=1}^{k/2-1} (time_{i+1} - time_i) - time_{1k}^{R/2R}$	$\beta i)^2$
views, Number of users who saw the original photo, until the kth reshare was posted, per unit time, or tients	0, k
$views_{0,k}^{\prime}$ Number of users who saw the original photo, until the k th reshare was posted, per unit time, or $views_{1,k-1,k}^{\prime}$ Number of users who saw the first $k-1$ reshares, until the k th reshare was posted, per unit time, or	

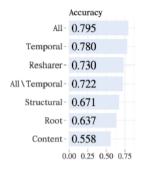


Figure 4: Using logistic regression, we are able to predict with near 80% accuracy whether the size of a cascade will reach the median (10) after observing the first k = 5 reshares.

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

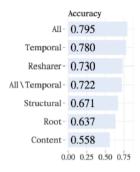


Figure 4: Using logistic regression, we are able to predict with near 80% accuracy whether the size of a cascade will reach the median (10) after observing the first k = 5 reshares.

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

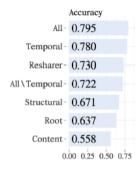


Figure 4: Using logistic regression, we are able to predict with near 80% accuracy whether the size of a cascade will reach the median (10) after observing the first k = 5 reshares.

- ► 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

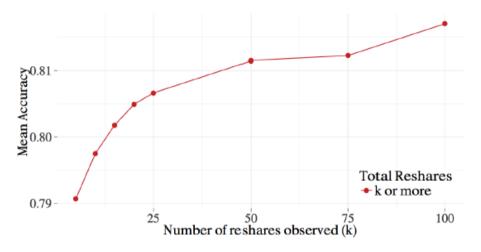
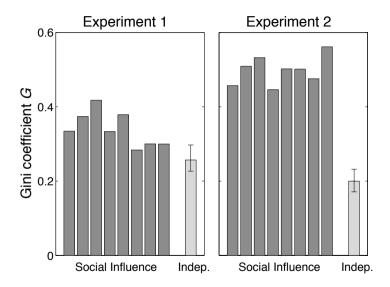


Figure 5: If we observe the first k reshares of a cascade, and want to predict whether the cascade will double in size, our prediction improves as we observe more of it.





gini coefficient: 0.787!



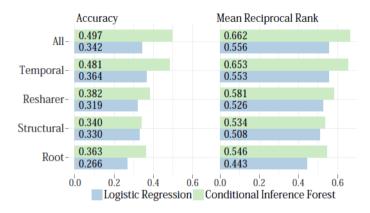


Figure 10: In predicting the largest cascade in clusters of 10 or more cascades of identical photos, we perform significantly above the baseline of 0.1.

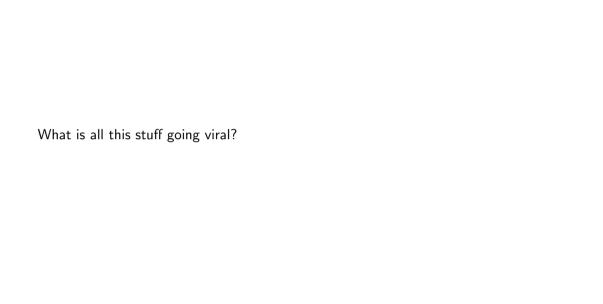
We can somewhat predict which of the identical seeds will spread, if we observe the beginning of each cascade

▶ asking the right question can be very important in research

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

- 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

- ▶ 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



- ▶ 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.