Can cascades be predicted?

Jure Leskovec (@jure)
Joint work with J. Cheng, H. Lakkaraju, J. McAuley, L. Adamic, A. Dow, J. Kleinberg
We are embedded in networks
Networks provide access to information!
We ...

- learn
- make decisions
- gain trust and safety

... based on the influence from our neighbors in the network

**For example:** Before purchasing electronics

- 50% of people do research online
- 68% of people consult friends [Burke]
Diffusion in Networks

- Networks provide a skeleton for the diffusion and flow of information!

More generally:

- **Contagion:** Behavior that spreads from a node to node like an epidemic
  - News, opinions, rumors
  - Word-of-mouth and product adoptions
  - Political mobilization
  - Infectious diseases
Diffusion in Networks

Contagions spread through the network like epidemics.
Contagions spread through the network like epidemics
Contagions spread through the network like epidemics
Diffusion in Networks

Contagions spread through the network like epidemics
Contagions spread through the network like epidemics.
Contagions spread through the network like epidemics
Contagions spread through the network like epidemics
Diffusion in Networks

Contagions spread through the network like epidemics
Diffusion in Networks

Contagions spread through the network like epidemics
An Information Cascade
Cascades in Viral Marketing

- People send and receive product recommendations, purchase products

5/30/2014  Jure Leskovec, Stanford University
Cascades in Viral Marketing

- People *send* and *receive* product recommendations, *purchase* products
People send and receive product recommendations, purchase products.
People **send and receive** product recommendations, **purchase** products.
People send and receive product recommendations, purchase products
People send and receive product recommendations, purchase products

Large online retailer: Jun ‘01 - May ‘03
- 16M recommendations on 500k products
- 4M customers
Recommendation Network

- Purchase following a recommendation
- Customer recommending a product
- Customer not buying a recommended product
As users of social media sites (re)share information, big cascades form.
Lada Adamic shared a link via Erik Johnston.

January 16, 2013

When life gives you an almost empty jar of nutella, add some ice cream...
(and other useful tips)

50 Life Hacks to Simplify your World
twistedsifter.com

Life hacks are little ways to make our lives easier. These low-budget tips and trick can help you organize and de-clutter space; prolong and preserve your products; or teach you...
Cascades form as people (re)share information with one another.
I fucking love science

Seriously. If you have a pizza with radius "z" and thickness "a", its volume is $V = \pi z^2 a$.

Lina von DelSienn, Iman Khallaf, 周明佳 and 73,191 others like this.

27,761 shares
How many people are using a particular hashtag?
Ma, Z., Sun, A., & Cong, G. (2013). On predicting the popularity of newly emerging hashtags in Twitter

Will a tweet get retweeted?

How does a large photo reshare cascade spread?

How will a cascade grow in the future?
Large cascades are extremely rare
How does Content Spread?

Same content, very different popularity
Increasing the strength of social influence increases both inequality and unpredictability of success.

Plan for the Talk

- Can cascades be predicted? [WWW ‘14]
  - Using complete Facebook data, can we predict which photos will get reshared a lot?

- Can cascades be created? [ICWSM ‘13]
  - Can we create machine generated submissions that get lots of activity (likes)?
Can cascades be predicted?

Cascades are predictable!

(*solution: cascade growth prediction problem)
Cascade prediction problem

How to formulate the prediction problem?
Formulating the problem

- Predict whether a cascade will get $>k$ reshares?
Predict whether a cascade will get \(>k\) reshares?

Most cascades are small. Class imbalance.
Formulating the Problem

- Predict the exact number of nodes in a cascade
Predict the exact number of nodes in a cascade

Cascade size distribution is heavily skewed. Outliers skew the error.
Only look at cascades with a minimum number of reshares and predict future growth
Only look at cascades with a minimum number of reshares and predict future growth.

Selection bias: Predicting over only a subset of data.
Will a cascade reach the median size?

For cascades of size $\geq k$ let the $f(k)$ be their median final size.

less than the median $f(k)$

more than the median $f(k)$
> Median? $\equiv$ Will it double?
Problem Formulation

- Will a cascade double in size?

\[ \leq 2k \text{ reshares} = f(k) \]

\[ > 2k \text{ reshares} = f(k) \]
Cascade Growth Prediction Problem

Given that a cascade has obtained $k$ reshares, will it grow beyond the median size $f(k)=2k$?
Given that a cascade has obtained $k$ reshares, will it grow beyond the median size $f(k) = 2k$?
We use anonymized Facebook photo resharing data from June 2013 to reconstruct reshare cascades using click, impression, and friend/follower data.

Using features of the cascade, we evaluate the performance of a classifier.

- 350m photos/day
- 150k photos ≥5 reshares
- 9m reshares total
Factors of Predictability

- **Content**
  (e.g. has overlaid text)

- **User**
  (e.g. follower count)

- **Structural**
  (e.g. proximity to root in $G$)

- **Temporal**
  (e.g. time between reshares)

5/30/2014  Jure Leskovec, Stanford University
Predictability of Cascades

Will a cascade double ($k=5$)?

- All: 0.795
- Temporal: 0.78
- User: 0.73
- All but temporal: 0.722
- Structural: 0.671
- Content: 0.558
How does predictability change with $k$?

- 5 reshares
- > 10 reshares?
- 20 reshares
- > 40 reshares?
How does predictability change with $k$?

- Two opposing arguments:
  - Easy: see more of the cascade
  - Hard: predict farther into the future
Predictability improves with $k$
Fix the minimum cascade size $R \geq k$

How does predictability change with $k$?

Is there a “sweet” spot when we have enough information to be able to predict?
Predictability improves with $k$

Accuracy

Number of reshares observed, $k$ (assuming minimum cascade size $R \geq 100$)
Network, Source & Content

- The original post (and poster) get less important with increasing $k$
Successful cascades get many views quickly, and achieve high conversion rates
Growth & the Initial Structure
Structure and Growth
Cascade Structure

- Predictability of cascade structure:
Wiener index: Mean all-pairs shortest path distance in a cascade

\[ d = 1.98 \]

\[ d = 2.47 \]

\[ d = 14.4 \]

Will a cascade have structural virality above the median?
Predictability of Structure

- Will a cascade have structural virality above the median?

Accuracy of 0.725

Temporal and structural features equally predictive
Increasing the strength of social influence increases both inequality and unpredictability of success.

Content and Popularity
Can we differentiate cascades of the same content?
Consider identical photos uploaded to Facebook, generating 983 clusters, 38k photos, 13m reshares

For each cluster, we select ten random cascades, and predict which was the largest

- Random guessing: 10% accuracy
Can we predict the largest of 10 random chosen cascades of the identical image?
Predictability

- Can we predict the largest of 10 random chosen cascades of the identical image?

Gini Coefficient of 0.787

Accuracy of 0.497

Mean Reciprocal Rank of 0.662
So...how do I make my posts go “viral”?
Maximize Content Success

Given a piece of content, can we maximize probability of its success?

Factors influencing popularity

- Community or Forum
- Time of posting
- Title of submission
- Popularity of user
- Previous submissions of same content

+ Content

and their confounding interplay!
How do we tease apart effects of various factors?

A dataset which accommodates:
- Resubmissions of same content
- Submissions across multiple communities
- Communities with varying characteristics
- Submissions by multiple users
Natural experiment: Every piece of content submitted multiple times

- 132K Reddit submissions
- 16.7K original submissions
- Average of 7 resubmissions per image

Data available at [http://snap.stanford.edu/data](http://snap.stanford.edu/data)
Content on Reddit

How well did this submission do?

score (upvotes - downvotes)

submission number (ordered by time)
Content on Reddit

REPOST SOMETHING FROM REDDIT ON REDDIT

GET MORE UPVOTES THAN ORIGINAL POST
Understanding Popularity

- Popularity
- Time
- User
- Community
- Title

I'm not sure I quite understand this piece
Submitted 2 years ago to pics by xxx
24 comments

Content
Our Approach

- Popularity = Community Model + Language Model
  
  - **Community model:** Choice of community + time of submission + previous submissions of same content
  
  - **Language model:** Linguistic features of submission title + language of community
Temporal Effects

![Average success over time graph](image)

- **Score (upvotes - downvotes)**
- **Time (UTC)**
- **Categories**: atheism, funny, gaming, gifs, pics, GifSound

5/30/2014

Jure Leskovec, Stanford University
Resubmissions are less popular (left), but can still be popular if we wait long enough (right)
Submissions won’t be successful in the same community twice (main diagonal)

Submissions won’t be successful if they already succeeded in a big community (low-rank structure)
The model is designed to account for five factors:

1. The inherent popularity of the content
2. The decay in popularity due to resubmitting the content
3. This decay should be discounted for old enough submissions
4. A penalty due to resubmitting to another community
5. A penalty due to resubmitting to the same community (also account for other factors, like the time of day, etc.)
Model: Language effects

Community specificity vs. success

Title originality vs. success

- Titles should match others in the same community, but should not be too similar
- Titles should differ from those previously used for the same content
Gramatical Structures

Parts of speech vs. success

- atheism
- funny
- gaming
- gifs
- pics

$p(A_{h,n} > 1)$

Nouns, Pronouns, Adjectives, Verbs, Adverbs, Interjections, Determiners

Part of speech
Performance on held-out test data:

<table>
<thead>
<tr>
<th>Model</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Community model only</td>
<td>0.528</td>
</tr>
<tr>
<td>Language model only</td>
<td>0.081</td>
</tr>
<tr>
<td>Community + language</td>
<td>0.618</td>
</tr>
</tbody>
</table>

Rating: upvotes – downvotes

- **Attention**: upvotes + downvotes: $R^2 = 0.58$
- **Engagement**: # comments: $R^2 = 0.68$
We generated **pairs** of 85 posts
We submitted them simultaneously to two different communities

‘**Good**’ posts got 3x as many upvotes
  - **Five** good titles reached the **front page** of their community
  - **Two** reached the front page of r/all
Good title: What I would do to someone I hate
- Votes: 7087/5228, Cmts.: 518

Why is this good?
- Original title
- Optimal length (not too short)
- POS tags: Interesting way of sentence structure compared to a flat tone syntax

Bad title: Funny gif
- Votes: 300/124, Cmts.: 9

Why is this bad?
- Not original, too generic (no specificity)
- Super short length
- Flat POS tag distribution
REPOST SOMETHING FROM REDDIT ON REDDIT

GET MORE UPVOTES THAN ORIGINAL POST
Cascades are predictable

- The cascade growth prediction problem, allows us to accurately predict the future growth of a cascade, even as it continues to grow in size

Cascades can be engineered
(to a large extent)
What have we learned about “viral” posts?

- Favor memes (i.e. post popular content)
- Have lots of followers (i.e. be popular)
- Know what your friends like, and what your friends’ friends like (i.e. be a marketing guru)
- Post at the right time
Further Qs: Interactions
Further Qs: Opinion dynamics

- Can this analysis help identify dynamics of polarization?

- Connections to mutation of information:
  - How does attitude and sentiment change in different parts of the network?
  - How does information change in different parts of the network?
Messages spreading through network require new ways of thinking about information dynamics and consumption

Feedback effects in networks:
- The feedback from using your social connections
  Some links are strengthened, others created:
    - Something that’s been going on for millennia
- The feedback from media that let you observe your place in the social network
  - A new and uncontrolled experiment
THANKS!
@jure
http://snap.stanford.edu
### References


