How Predictable is Information Diffusion?

Travis Martin, Jake Hofman, Amit Sharma, Ashton Anderson, and Duncan Watts
How far will this spread?

Neil deGrasse Tyson
@neiltyson


12:48 PM - 13 Feb 2016
How far will this spread?

Neil deGrasse Tyson
@neiltyson


12:48 PM - 13 Feb 2016

RETWEETS 21,984 LIKES 35,477
Why is so difficult to predict success?

Do we need bigger data and better models?


Or is information diffusion inherently unpredictable?
Outline

• Understanding diffusion: What we know and how we got here

• Predicting success: Evaluating the state-of-the-art under a unified framework

• Theoretical limits: Exploring the limits to predicting success
Understanding Diffusion

(What we know and how we got here)
~1950s: Small-scale surveys of individual interactions
~1950s: Small-scale surveys of individual interactions

Table 56—The "Cosmopolitans" Among the Opinion Leaders Are in Fashions and Public Affairs

<table>
<thead>
<tr>
<th></th>
<th>Marketing Non-Leaders</th>
<th>Fashion Non-Leaders</th>
<th>Public Affairs Non-Leaders</th>
<th>Movie Non-Leaders</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low Ed'n</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mass media</td>
<td>27%</td>
<td>39%</td>
<td>50%</td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>(88)</td>
<td>(79)</td>
<td>(30)</td>
<td>(64)</td>
</tr>
<tr>
<td></td>
<td>100%</td>
<td>(324)</td>
<td>(330)</td>
<td>(381)</td>
</tr>
<tr>
<td></td>
<td>High Ed'n</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>48%</td>
<td>53%</td>
<td>55%</td>
<td>45%</td>
</tr>
<tr>
<td></td>
<td>(77)</td>
<td>(81)</td>
<td>(51)</td>
<td>(58)</td>
</tr>
<tr>
<td></td>
<td>100%</td>
<td>(219)</td>
<td>(218)</td>
<td>(148)</td>
</tr>
</tbody>
</table>

Katz & Lazarsfeld (1955)
~1960s: Mathematical models of aggregate adoption

Rogers (1962), Bass (1969)
\( p > \frac{(1 + \epsilon) \ln n}{n} \)

Erdős & Rényi (1959)
~1990s: Empirical structure and dynamics of networks

~2000s: Empirical analyses of large-scale diffusion events

Liben-Nowell & Kleinberg (2007)
~2010s: Characterizing online information flows

Wu, Hofman, Mason, Watts (2011)
~2010s: Cataloging empirical diffusion structures

Fig. 2. The distribution of diffusion cascade structures unreciprocated and public. Whereas Zync, Friendsense, and Yahoo! Voice clearly exhibit positive externalities in the sense that the utility of the products in question increases with the number of adopting neighbors, such network effects are less likely in the remaining domains. And whereas the diffusion cascades on Twitter generally terminated within a day or two, the Secretary Game and Friendsense spread actively for several weeks, while cascades on Yahoo! Voice extended over several years.

Given the heterogeneity in data collection, timescales (ranging from days to years), and the nature of adoptions described above, the distribution of diffusion structures across all seven cases is striking in its similarity. Fig. 2A shows the frequency of cascades accounted for by the most commonly occurring tree structures across the seven domains we study. The vast majority of instances—ranging from 73% to 95% across domains—show no diffusion at all (i.e., the tree consists only of the seed), while the next most frequent outcome is in all cases a single additional adopter. In fact, the same seven simple tree structures account for upwards of 97% of cascades in each domain. Figs. 2B and 2C complement this result, showing that the distributions of tree size and depth, respectively, are likewise extremely skewed. In all domains, less than 1% of cascades consist of more than seven nodes, and less than 4% extend further than one degree from the seed node.

Although the similarity across domains is striking, our finding that most cascades are small and shallow is not, on its own, surprising. A number of recent empirical studies of online diffusion [Adar and Adamic 2005; Leskovec et al. 2007; Bakshy et al. 2009; Sun et al. 2009; Bakshy et al. 2011] have also observed that the size distribution of diffusion events is right-skewed and heavy-tailed, which necessarily implies that most events are small; indeed, Leskovec et. al [Leskovec et al. 2007] even identify many of the same motifs. The usual intuition regarding heavy-tailed distributions, however, is that large events, although rare, are sufficiently large to dominate certain key proper-

Goel, Goldstein, Watts (2012)
~2010s: Cataloging empirical diffusion structures

• There is a striking concentration of attention online, in support of the two-step flow of information

• Most things don’t spread, but when they do, there is a great deal of diversity in diffusion patterns

• There is almost no correlation between how things diffuse and how far they spread

• Existing diffusion models fail to account for this diversity in outcomes
Predicting Success

(Evaluating the state-of-the-art under a unified framework)
Background: Predicting the success of diffusion events

Bakshy, Hofman, Mason, Watts (2011)

- Looked at 75M diffusion events across 1M users
- Found a relatively low correlation ($R^2 \sim 30\%$) between predicted and actual cascade sizes
- Almost all predictive power comes from examining past performance of a user or piece of content
Background: Predicting the success of diffusion events
Bakshy, Hofman, Mason, Watts (2011)

- Looked at 75M diffusion events across 1M users
- Found a relatively low correlation ($R^2 \sim 30\%$) between predicted and actual cascade sizes
- Almost all predictive power comes from examining past performance of a user or piece of content

How much better can we do?
Related work

• Hong & Davidson (2010): Will a given user be retweeted? Topic model features outperform baselines ($F1 = 0.47$)

• Petrovic et. al. (2011): Will a given tweet be retweeted? Social and content features beat humans ($F1 = 0.46$)

• Jenders et. al. (2013): Will a cascade reach a minimum size? Content features lead to good performance ($F1 = 0.90$)

• Tan et. al. (2014): Which of two tweets will spread further? Detailed wording features are informative ($\text{Accuracy} = 0.65$)

• Cheng et. al. (2014): Will a cascade double in size? Temporal features provide good performance ($\text{AUC} = 0.88$)
All of this work examines a different question with a different measure of success, evaluated on a different subset of data, making it difficult to assess overall progress\(^1\)

\(^1\)http://hunch.net/?p=22
Ex-ante prediction

We focus on predictions made prior to events of interest

“X will succeed because of properties A, B, and C”

vs.

“X will succeed tomorrow because it is successful today”
A unified framework: Luck vs. skill

- Model success $S$ as a mix of skill $Q$ and luck $\epsilon$:
  $$S = f(Q) + \epsilon$$

- Measure the fraction of variance remaining after conditioning on skill:
  $$F = \frac{\mathbb{E}[\text{Var}(S|Q)]}{\text{Var}(S)} = 1 - R^2$$

- $R^2 = 1$ in a pure skill world, $R^2 = 0$ in pure luck world

---

$^2$Formalizes Maboussin (2012)
Data

- Examined all 1.4B tweets containing URLs posted in February 2015
Data

- Examined all 1.4B tweets containing URLs posted in February 2015
- Eliminated spam using internal Microsoft classifier
- Restricted attention to tweets containing URLs from the top 100 English-speaking domains with the most unique adopters
- Resulted in 850M tweets from 50M distinct users covering news, entertainment, videos, images, and products
- Measured the total cascade size for each seed tweet
Data

- Examined all 1.4B tweets containing URLs posted in February 2015
- Eliminated spam using internal Microsoft classifier
- Restricted attention to tweets containing URLs from the top 100 English-speaking domains with the most unique adopters

Resulted in 850M tweets from 50M distinct users covering news, entertainment, videos, images, and products.

Measured the total cascade size for each seed tweet.
Data

- Examined all 1.4B tweets containing URLs posted in February 2015
- Eliminated spam using internal Microsoft classifier
- Restricted attention to tweets containing URLs from the top 100 English-speaking domains with the most unique adopters
- Resulted in 850M tweets from 50M distinct users covering news, entertainment, videos, images, and products
Data

- Examined all 1.4B tweets containing URLs posted in February 2015
- Eliminated spam using internal Microsoft classifier
- Restricted attention to tweets containing URLs from the top 100 English-speaking domains with the most unique adopters
- Resulted in 850M tweets from 50M distinct users covering news, entertainment, videos, images, and products
- Measured the total cascade size for each seed tweet
Most users in our dataset have relatively few followers, although low-degree users are under-represented.
Cascade sizes

Most cascades are small, fewer than 3% reach 10 or more users
Activity by degree

Most cascades are started by low-degree users
Cascade size by degree

Cascades initiated by high-degree users tend to have larger reach

![Diagram showing the relationship between number of followers of a user and mean cascade size for a typical user. The x-axis represents the number of followers, ranging from 10 to 10,000,000, and the y-axis represents the mean cascade size, ranging from 0.1 to 100,000. The graph shows a logarithmic scale for both axes. The black line represents the trend, with shaded areas indicating the variability. There are markers for different numbers of users: 1, 100, 10,000, and 1,000,000.]
Used a random forest to estimate success (cascade size) given skill (available features)

- **Basic content features**: URL domain, time of tweet, spam score, ODP category
- **Basic user features**: number of followers, number of friends, number of posts, account creation time
- **Topic features**: the most probable Latent Dirichlet Allocation topic for each user and tweet, along with an interaction term
- **Past success**: the average number of retweets received by each URL and user in the past
Predictive performance

Our best model explains roughly half of the variance in outcomes

<table>
<thead>
<tr>
<th>Model</th>
<th>Tweet time</th>
<th>Domain</th>
<th>Spam score</th>
<th>Category</th>
<th>Tweet topic</th>
<th>Past url success</th>
<th>User time</th>
<th>Followers</th>
<th>Friends</th>
<th>_statuses</th>
<th>User topic</th>
<th>Past user success</th>
<th>Topic interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Basic content</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Content, topic</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Content, past succ.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Basic user</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. User, topic</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. User, past succ.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Content, user</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. All</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Predictive performance

Content features alone perform poorly

<table>
<thead>
<tr>
<th>Model</th>
<th>Tweet time</th>
<th>Domain</th>
<th>Spam score</th>
<th>Category</th>
<th>Tweet topic</th>
<th>Past url success</th>
<th>User time</th>
<th>Followers</th>
<th>Statuses</th>
<th>User topic</th>
<th>Past user success</th>
<th>Topic interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Basic content</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>2. Content, topic</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>3. Content, past succ.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>4. Basic user</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>5. User, topic</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>6. User, past succ.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>7. Content, user</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>8. All</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Graph showing $R^2$ values for different models.
Predictive performance

Basic user features provide a reasonable boost in performance

<table>
<thead>
<tr>
<th>Model</th>
<th>Tweet time</th>
<th>Domain</th>
<th>Spam score</th>
<th>Category</th>
<th>Tweet topic</th>
<th>Past url success</th>
<th>User time</th>
<th>Followers</th>
<th>Friends</th>
<th>Statuses</th>
<th>User topic</th>
<th>Past user success</th>
<th>Topic interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Basic content</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>2. Content, topic</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>3. Content, past succ.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>4. Basic user</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>5. User, topic</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>6. User, past succ.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>7. Content, user</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>8. All</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
Predictive performance

Past user success alone accounts for almost all of predictive power

<table>
<thead>
<tr>
<th>Model</th>
<th>Tweet time</th>
<th>Domain</th>
<th>Spam score</th>
<th>Category</th>
<th>Tweet topic</th>
<th>Past url success</th>
<th>User time</th>
<th>Followers</th>
<th>Friends</th>
<th>Statuses</th>
<th>User topic</th>
<th>Past user success</th>
<th>Topic interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Basic content</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>2. Content, topic</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>3. Content, past succ.</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>4. Basic user</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>5. User, topic</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>6. User, past succ.</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>7. Content, user</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>8. All</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
</tbody>
</table>
Summary of empirical results

- This is the **best known model** since Bakshy et. al., boosting performance from $R^2 \sim 30\%$ to $R^2 \sim 50\%$

- Both models derive their **predictive power** from the same simple feature: a user’s **past success**

- **Content features** are only **weakly informative**

- **Performance plateaus** as we add more features, suggesting a possible limit to the **predictability** of diffusion outcomes
Theoretical limits

(Exploring the limits to predicting success)
Simulations

• In practice we can never rule out missing features or superior models, so we turn to numerical simulations where we have full access to and control of all relevant information.

• Looked at the variation in outcomes when we repeatedly seed the same user with the same content.

• Examined how this varies with content heterogeneity and estimation error.
Simulations

- Created a scale-free network similar to Twitter but smaller in size
- Simulated 8B cascades using a standard SIR model
- Initiated 1,000 cascades for each combination of 10,000 different seed users and 800 different infectiousness values
- Carefully matched distributions of user activity and cascade size to our empirical data
Repeatedly seed the same user with the same content

Outcomes are highly predictable when all content is identical

Perfect knowledge of identical content

Content heterogeneity
Theoretical limit on predictability
Average content quality

- $R_0^* = 0.10$
- $R_0^* = 0.20$
- $R_0^* = 0.30$
- $R_0^* = 0.40$

How Predictable is Information Diffusion?
Repeatedly seed the same user with the same content

Predictive performance decreases sharply with content diversity (e.g., a 15% variation around $R_0^* = 0.2$ gives an $R^2$ of 60%)

Perfect knowledge of diverse content
Repeatedly seed the same user with the same content

Outcomes are highly predictable assuming exact quality estimates

Theoretical limit on predictability

Perfect knowledge of identical content

Average content quality

- $R_0 = 0.10$
- $R_0 = 0.20$
- $R_0 = 0.30$
- $R_0 = 0.40$

Error in estimating quality

How Predictable is Information Diffusion?
Repeatedly seed the same user with the same content

Predictive performance decreases sharply with estimation error (e.g., $R^2 < 60\%$ with 30\% error in estimating $R_0^* = 0.3$)
Summary of theoretical results

- Our simulations that suggest that it is the diffusion process itself that is unpredictable, rather than our ability to estimate or model it.

- Predictability decreases sharply with content diversity.

- Likewise, small errors in estimating quality severely limit predictability.

- We emphasize the qualitative nature of these results and the approach to assessing predictability, rather than the specific numerical outcomes presented here.
Conclusions
Conclusions

Most things don't spread, but when they do, it’s difficult to predict success.
Conclusions

Despite a great deal of research on the topic, it’s difficult to assess long-term progress in predicting success.
State-of-the-art models explain roughly half of the variance in outcomes, based primarily on past success.
This is likely due to randomness in diffusion process itself, rather than our ability to estimate or model it.