

How Predictable is the Spread of Information?

Jake Hofman

Microsoft Research NYC

August 17, 2018

How far will this spread?



Neil deGrasse Tyson

@neiltyson



Follow

1916: Einstein predicts Gravity Waves. 1917: He lays the foundation for Lasers. 2016: Gravity Waves discovered using Lasers.

RETWEETS

???

LIKES

???

12:48 PM - 13 Feb 2016

How far will this spread?

 Neil deGrasse Tyson 
@neiltyson

1916: Einstein predicts Gravity Waves. 1917: He lays the foundation for Lasers. 2016: Gravity Waves discovered using Lasers.

RETWEETS **21,984** LIKES **35,477**



12:48 PM - 13 Feb 2016

Why is so difficult to predict success?

Do we need bigger data and better models?



Neil deGrasse Tyson

@neiltyson



Follow

1916: Einstein predicts Gravity Waves. 1917: He lays the foundation for Lasers. 2016: Gravity Waves discovered using Lasers.

RETWEETS

21,984

LIKES

35,477



12:48 PM - 13 Feb 2016

Or is information diffusion inherently unpredictable?

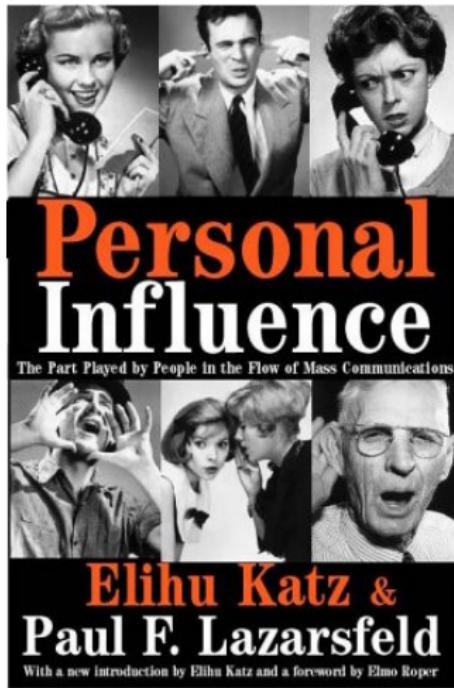
Outline

- **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

Two-step flow model

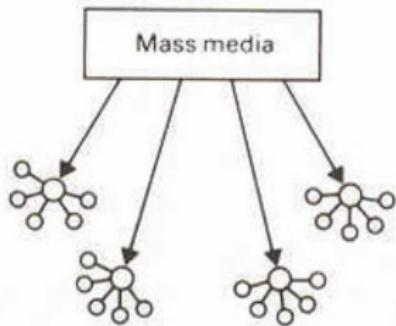
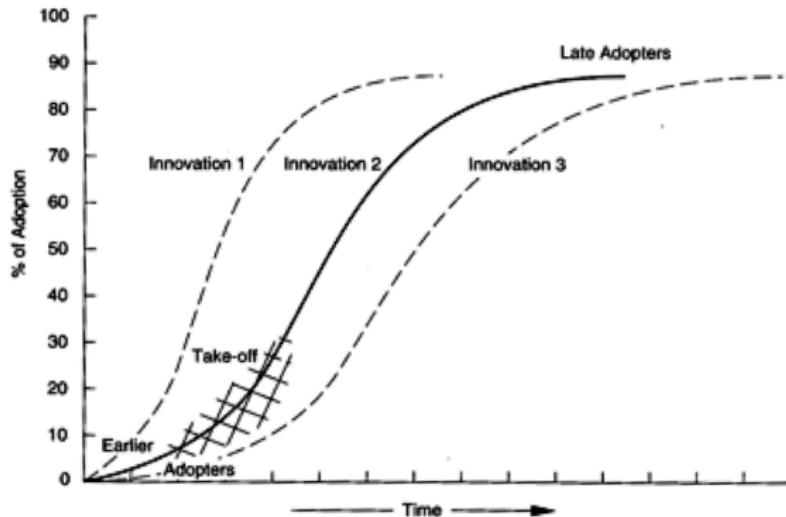


Table 56—The "Cosmopolitans" Among the Opinion Leaders Are in Fashions and Public Affairs

| | PER CENT WHO READ BOTH OUT-OF-TOWN NEWSPAPERS AND NEWS IN NATIONAL MAGAZINES | | | | | | | |
|-----------|--|-------------|---------|-------------|----------------|-------------|---------|-------------|
| | Marketing | | Fashion | | Public Affairs | | Movie | |
| | Leaders | Non-Leaders | Leaders | Non-Leaders | Leaders | Non-Leaders | Leaders | Non-Leaders |
| Low Ed'n | 27% | 20% | 39% | 17% | 50% | 20% | 25% | 24% |
| 100% = | (88) | (324) | (79) | (330) | (30) | (381) | (64) | (159) |
| High Ed'n | 48% | 43% | 53% | 41% | 55% | 41% | 45% | 47% |
| 100% = | (77) | (219) | (81) | (218) | (51) | (247) | (58) | (148) |

Katz & Lazarsfeld (1955)

~1960s: Mathematical models of aggregate adoption



Rogers (1962), Bass (1969)

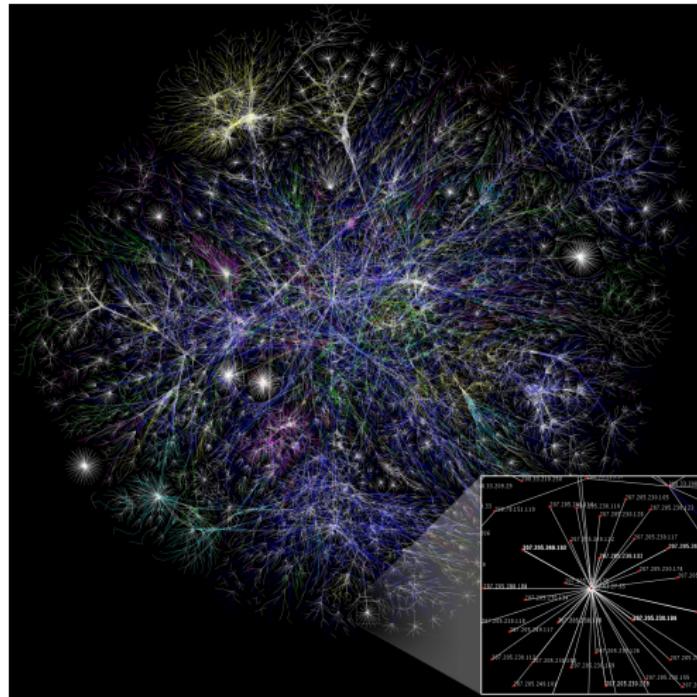
~1960s: Random graph theory



$$p > \frac{(1 + \epsilon) \ln n}{n}$$

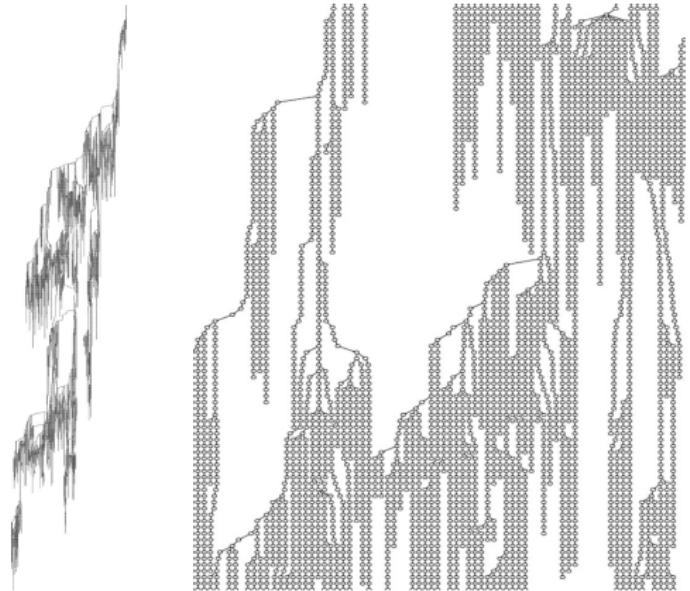
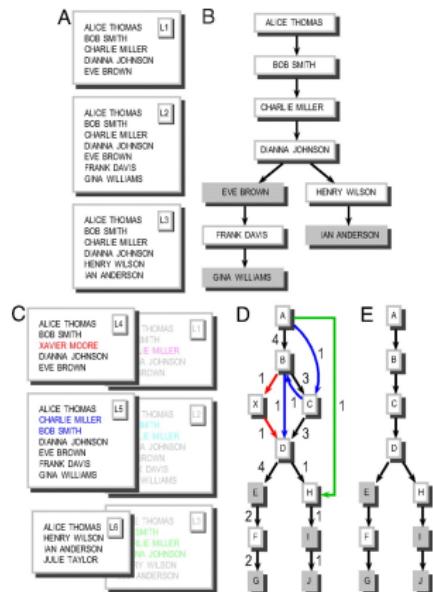
Erdős & Rényi (1959)

~1990s: Empirical structure and dynamics of networks



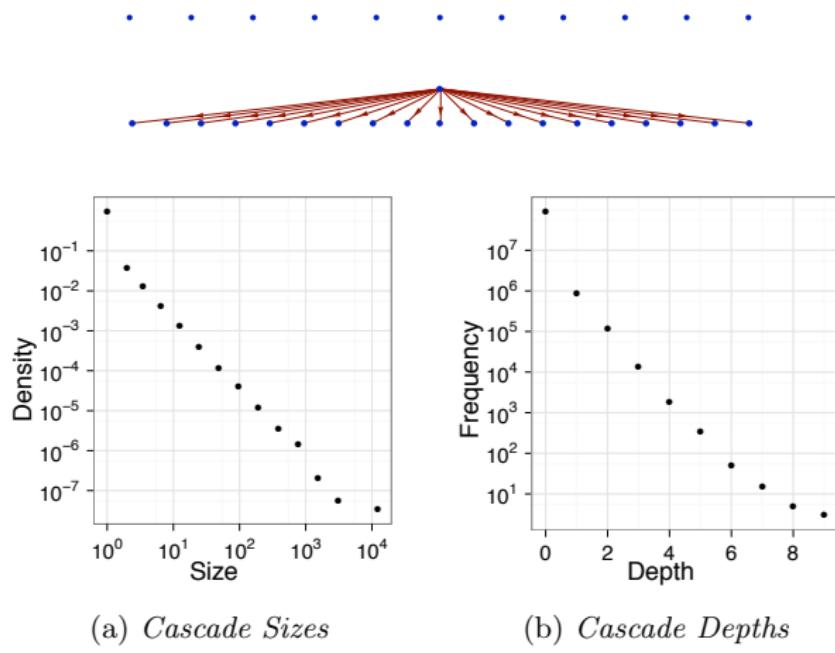
Newman, Barabasi, Watts (2006)

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



Liben-Nowell & Kleinberg (2007)

~2010s: Cataloging empirical diffusion structures

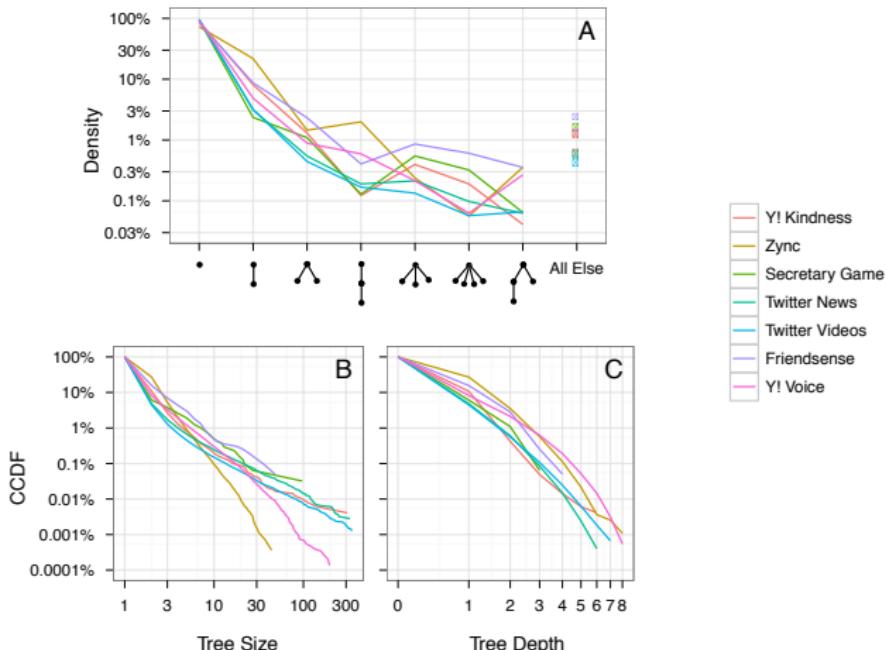


(a) *Cascade Sizes*

(b) *Cascade Depths*

Bakshy, Hofman, Mason, Watts (2011)

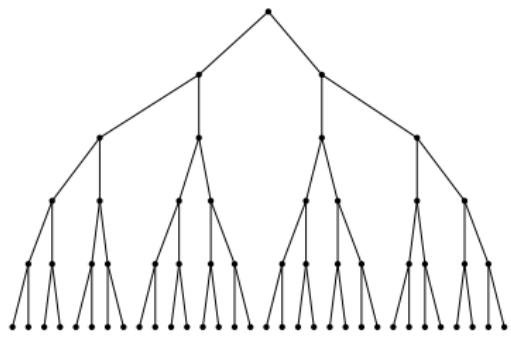
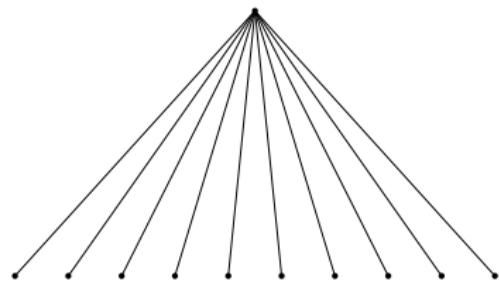
~2010s: Cataloging empirical diffusion structures



Goel, Goldstein, Watts (2012)

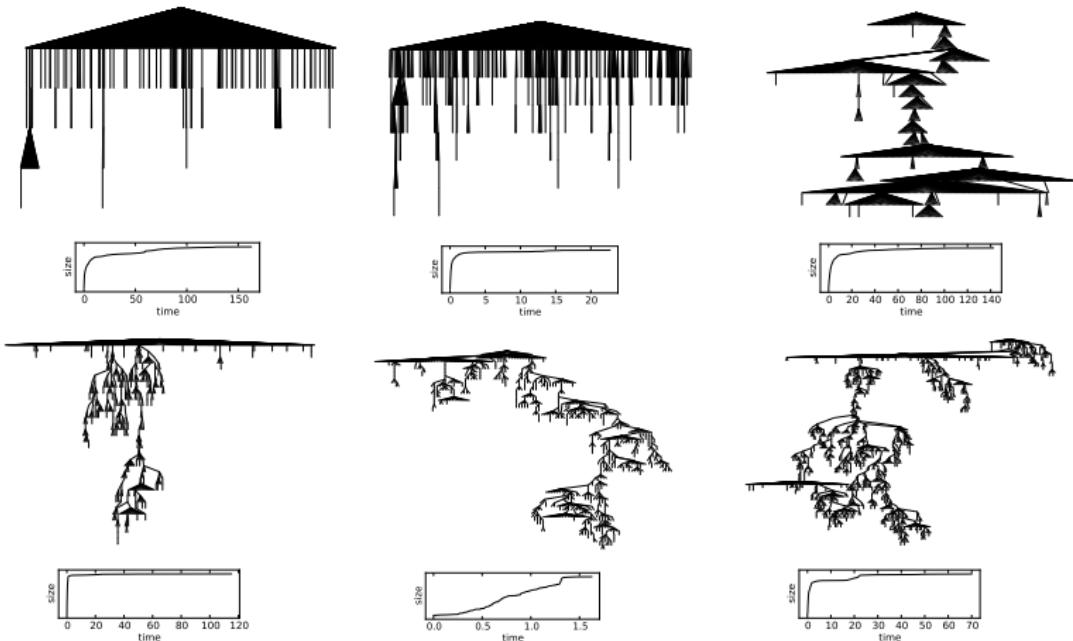
~2010s: Cataloging empirical diffusion structures

How do the popular things become popular?



Goel, Anderson, Hofman, Watts (2015)

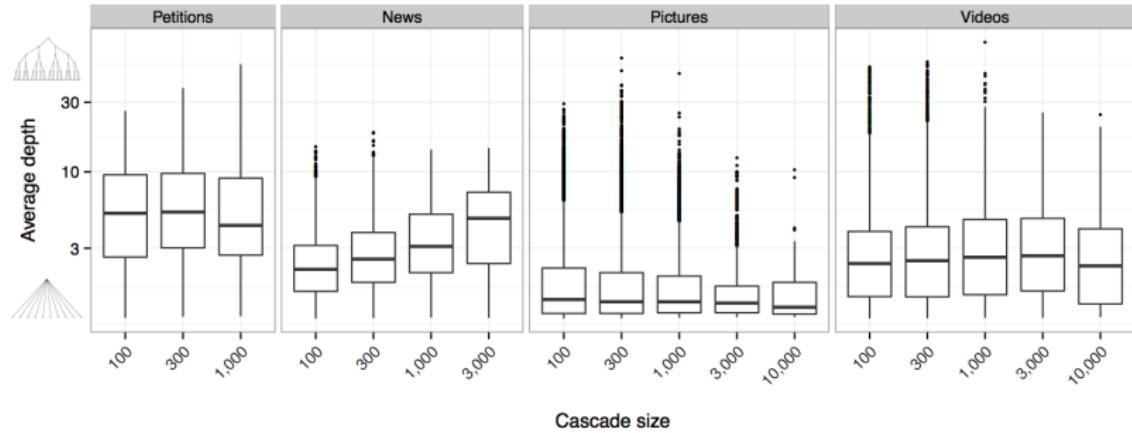
~2010s: Cataloging empirical diffusion structures



Goel, Anderson, Hofman, Watts (2015)

~2010s: Cataloging empirical diffusion structures

Size is relatively poor predictive of structure



Goel, Anderson, Hofman, Watts (2015)

- 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

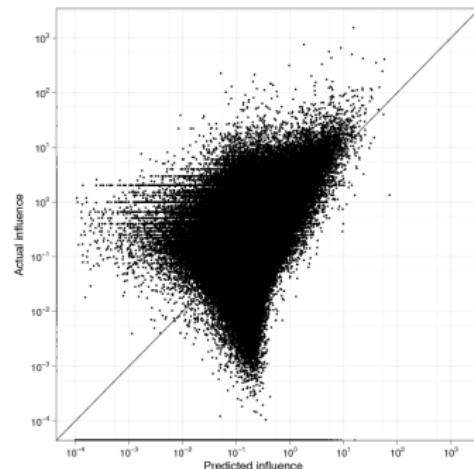
Predicting Success

(Evaluating the state-of-the-art under a unified framework)

Background: Predicting the success of diffusion events

Bakshy, Hofman, Mason, Watts (2011)

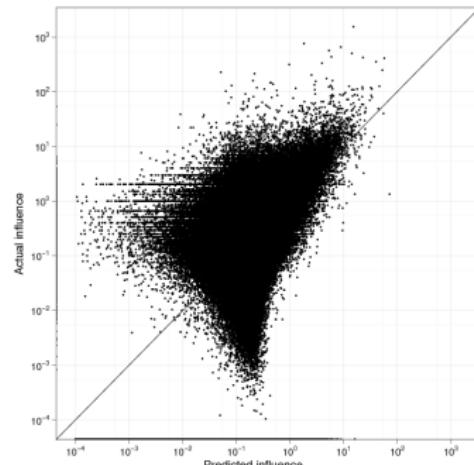
- Looked at 75M diffusion events across 1M users
- Found a relatively **low correlation** ($R^2 \sim 30\%$) between predicted and actual cascade sizes
- Almost all predictive power comes from examining **past performance** of a user or piece of content



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 (Accuracy = 0.65)
- Cheng et. al. (2014): Will a cascade double in size?
Temporal features provide good performance ($AUC = 0.88$)

Progress?

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>

Progress?

Same data, same model ... seemingly different answers

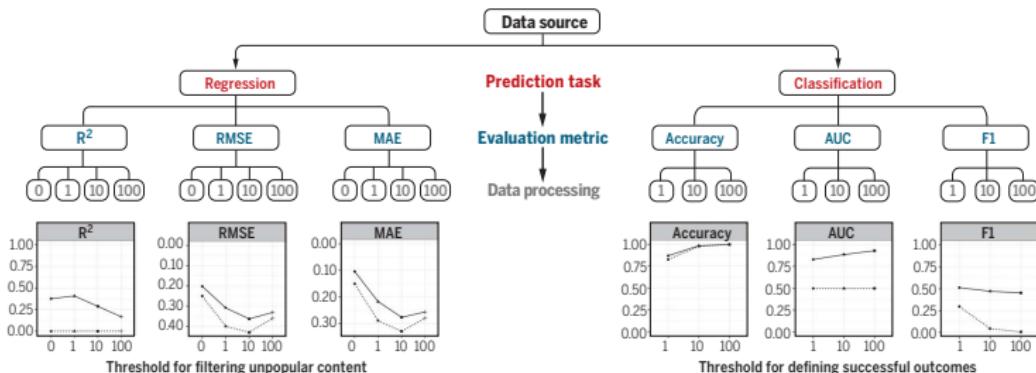


Fig. 1. A single question may correspond to many research designs, each yielding different answers. (Top) A depiction of the many choices involved in translating the problem of understanding diffusion cascades into a concrete prediction task, including the choice of data source, task, evaluation metric, and data preprocessing. The preprocessing choices shown at the terminal nodes refer to the threshold used to filter observations for regression or define successful outcomes for classification. Cascade sizes were log-transformed for all of the regression tasks. (Bottom) The results of each prediction task, for each

metric, as a function of the threshold used in each task. The lower limit of each vertical axis gives the worst possible performance on each metric, and the top gives the best. Dashed lines represent the performance of a naive predictor (always forecasting the global mean for regression or the positive class for classification), and solid lines show the performance of the fitted model. R^2 , coefficient of determination; AUC, area under the ROC curve; RMSE, root mean squared error; MAE, mean absolute error; F1 score, the harmonic mean of precision and recall.

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. skill¹

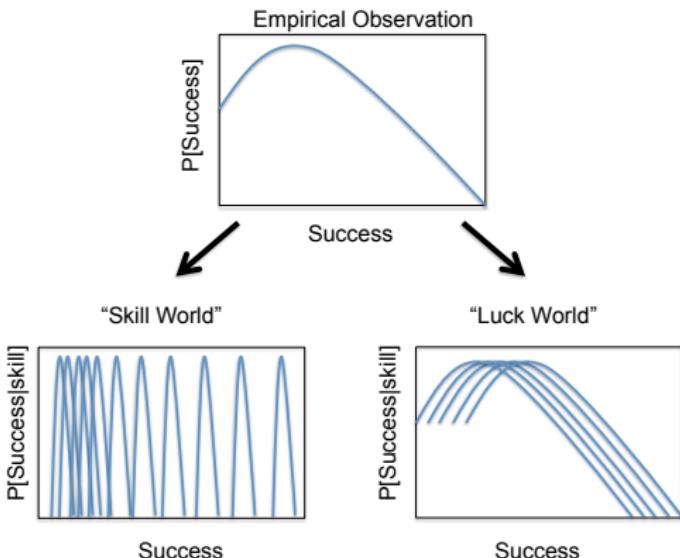
- Model success S as a mix of skill Q and luck ϵ :

$$S = f(Q) + \epsilon$$

- 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



¹Formalizes 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

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

Data

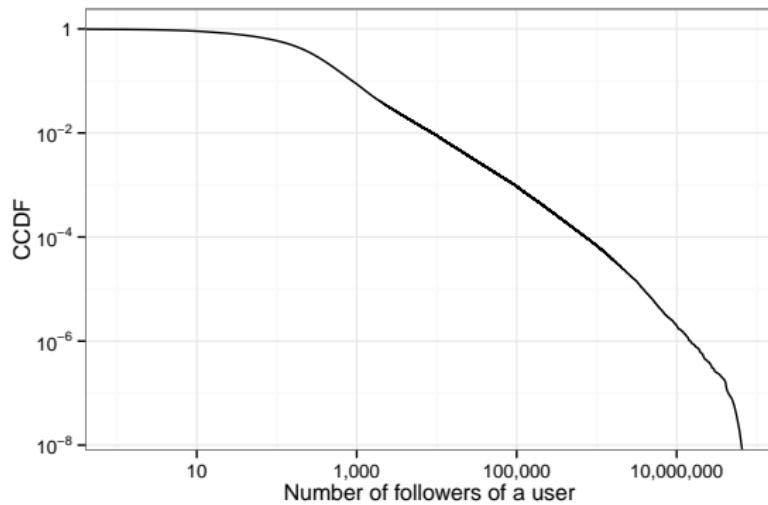
- Examined all 1.4B tweets containing URLs posted in February 2015
- Eliminated spam using internal Microsoft classifier
- Restricted attention to tweets containing URLs from the top 100 English-speaking domains with the most unique adopters
- Resulted in 850M tweets from 50M distinct users covering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

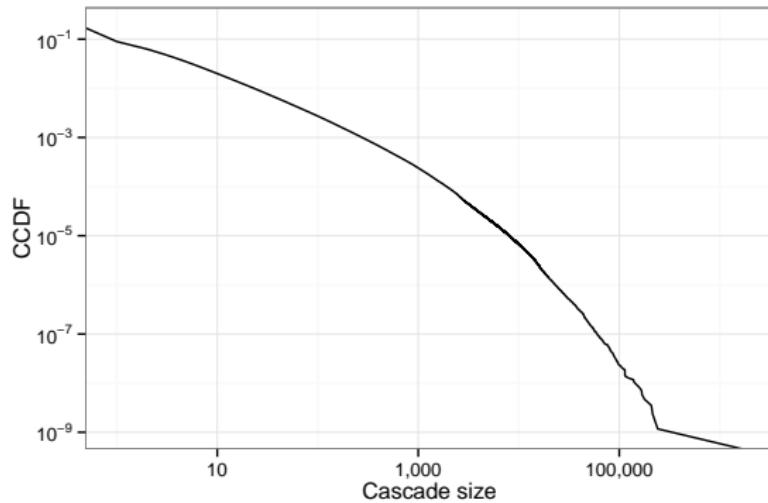
User distribution

Most users in our dataset have relatively few followers, although low-degree users are under-represented



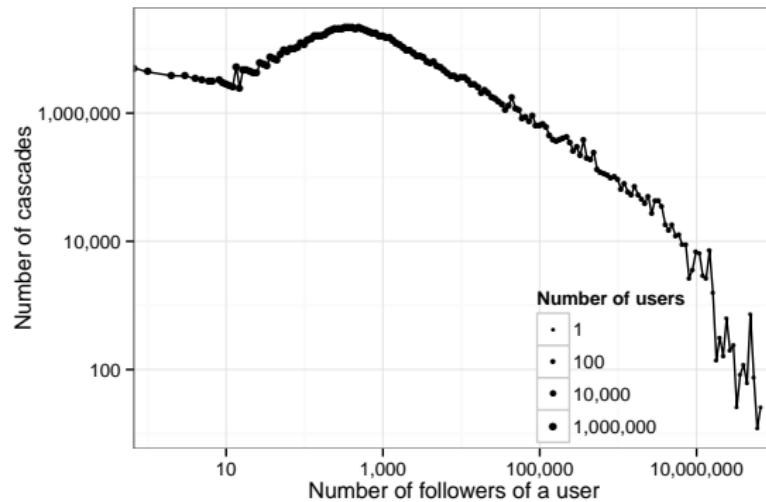
Cascade sizes

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



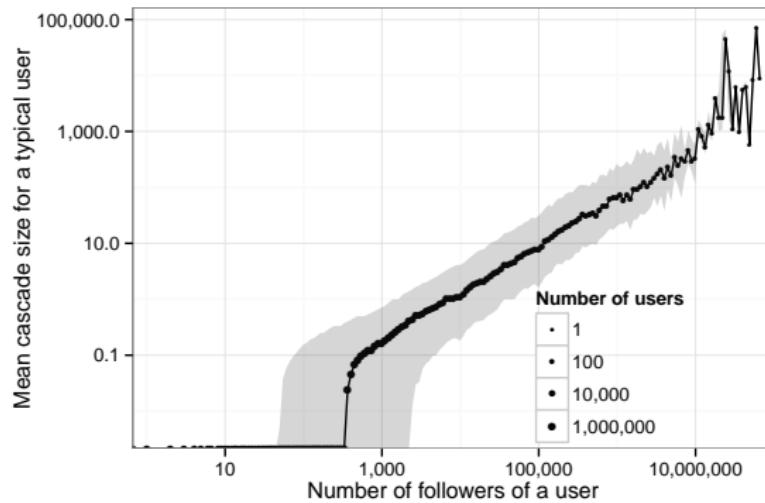
Activity by degree

Most cascades are started by low-degree users



Cascade size by degree

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



Predictive features

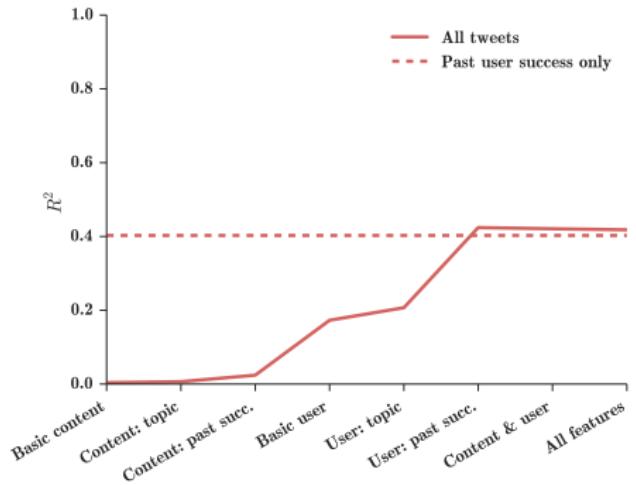
Used a random forest to estimate success (cascade size)
given skill (available features)

- **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 40% of the variance in outcomes

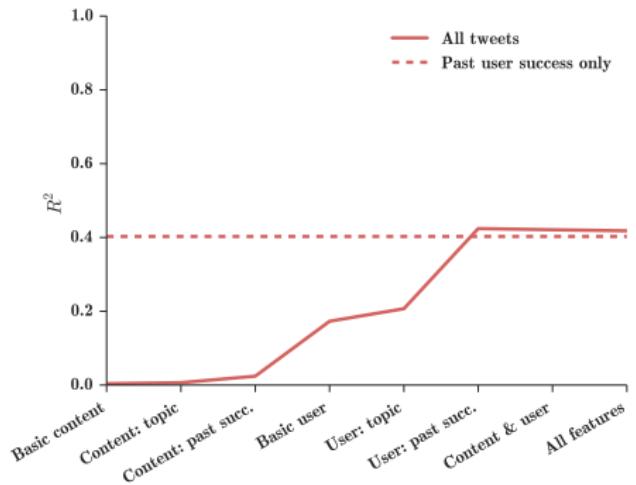
| Model | <i>Tweet time</i> | <i>Domain</i> | <i>Spam score</i> | <i>Category</i> | <i>Tweet topic</i> | <i>Past url success</i> | <i>User time</i> | <i>Followers</i> | <i>Friends</i> | <i>Statuses</i> | <i>User topic</i> | <i>Past user success</i> | <i>Topic interaction</i> |
|------------------------|-------------------|---------------|-------------------|-----------------|--------------------|-------------------------|------------------|------------------|----------------|-----------------|-------------------|--------------------------|--------------------------|
| 1. Basic content | ✓ | ✓ | ✓ | ✓ | | | | | | | | | |
| 2. Content, topic | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | | | | |
| 3. Content, past succ. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | | | |
| 4. Basic user | | | | | | ✓ | ✓ | ✓ | ✓ | ✓ | | | |
| 5. User, topic | | | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | |
| 6. User, past succ. | | | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| 7. Content, user | | | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| 8. All | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |



Predictive performance

Content features alone perform poorly

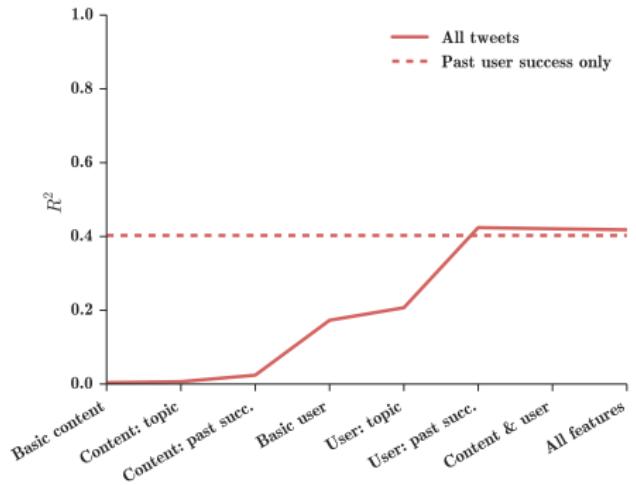
| Model | <i>Tweet time</i> | <i>Domain</i> | <i>Spam score</i> | <i>Category</i> | <i>Tweet topic</i> | <i>Past url success</i> | <i>User time</i> | <i>Followers</i> | <i>Friends</i> | <i>Statuses</i> | <i>User topic</i> | <i>Past user success</i> | <i>Topic interaction</i> |
|------------------------|-------------------|---------------|-------------------|-----------------|--------------------|-------------------------|------------------|------------------|----------------|-----------------|-------------------|--------------------------|--------------------------|
| 1. Basic content | ✓ | ✓ | ✓ | ✓ | | | | | | | | | |
| 2. Content, topic | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | | | | |
| 3. Content, past succ. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | | | |
| 4. Basic user | | | | | | ✓ | ✓ | ✓ | ✓ | | | | |
| 5. User, topic | | | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | |
| 6. User, past succ. | | | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| 7. Content, user | | | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| 8. All | | | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |



Predictive performance

Basic user features provide a reasonable boost in performance

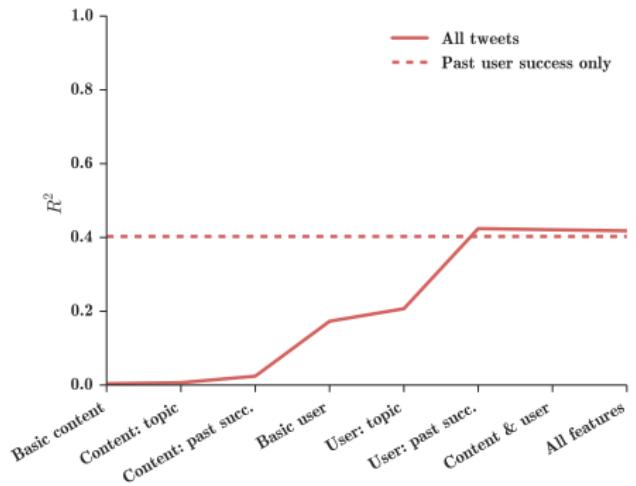
| Model | <i>Tweet time</i> | <i>Domain</i> | <i>Spam score</i> | <i>Category</i> | <i>Tweet topic</i> | <i>Past url success</i> | <i>User time</i> | <i>Followers</i> | <i>Friends</i> | <i>Statuses</i> | <i>User topic</i> | <i>Past user success</i> | <i>Topic interaction</i> |
|------------------------|-------------------|---------------|-------------------|-----------------|--------------------|-------------------------|------------------|------------------|----------------|-----------------|-------------------|--------------------------|--------------------------|
| 1. Basic content | ✓ | ✓ | ✓ | ✓ | | | | | | | | | |
| 2. Content, topic | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | | | | |
| 3. Content, past succ. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | | | |
| 4. Basic user | | | | | | ✓ | ✓ | ✓ | ✓ | | | | |
| 5. User, topic | | | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | |
| 6. User, past succ. | | | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| 7. Content, user | | | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| 8. All | | | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |



Predictive performance

Past user success alone accounts for almost all of predictive power

| Model | <i>Tweet time</i> | <i>Domain</i> | <i>Spam score</i> | <i>Category</i> | <i>Tweet topic</i> | <i>Past url success</i> | <i>User time</i> | <i>Followers</i> | <i>Friends</i> | <i>Statuses</i> | <i>User topic</i> | <i>Past user success</i> | <i>Topic interaction</i> |
|------------------------|-------------------|---------------|-------------------|-----------------|--------------------|-------------------------|------------------|------------------|----------------|-----------------|-------------------|--------------------------|--------------------------|
| 1. Basic content | ✓ | ✓ | ✓ | ✓ | | | | | | | | | |
| 2. Content, topic | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | | | | |
| 3. Content, past succ. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | | | |
| 4. Basic user | | | | | | ✓ | ✓ | ✓ | ✓ | ✓ | | | |
| 5. User, topic | | | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | |
| 6. User, past succ. | | | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| 7. Content, user | | | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| 8. All | | | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |



Summary of empirical results

- This is the **best known model** since Bakshy et. al., boosting performance from $R^2 \sim 30\%$ to $R^2 \sim 40\%$
- 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

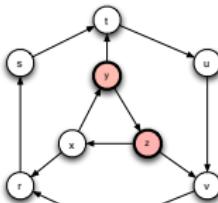
(Exploring the limits to predicting success)

Simulations

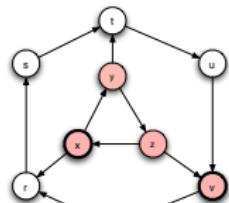
- In practice we can never rule out missing features or superior models, so we turn to numerical simulations where we have full access to and control of all relevant information
- Looked at the variation in outcomes when we repeatedly seed the same user with the same content
- Examined how this varies with content heterogeneity and estimation error

Simulations

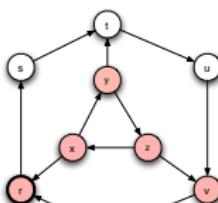
- Created a scale-free network similar to Twitter but smaller in size
- Simulated 8B cascades using a standard SI model
- Initiated 1,000 cascades for each combination of 10,000 different seed users and 800 different infectiousness values
- Carefully matched distributions of user activity and cascade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



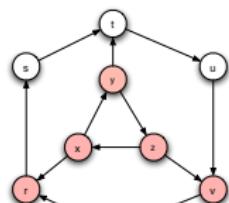
(a)



(b)



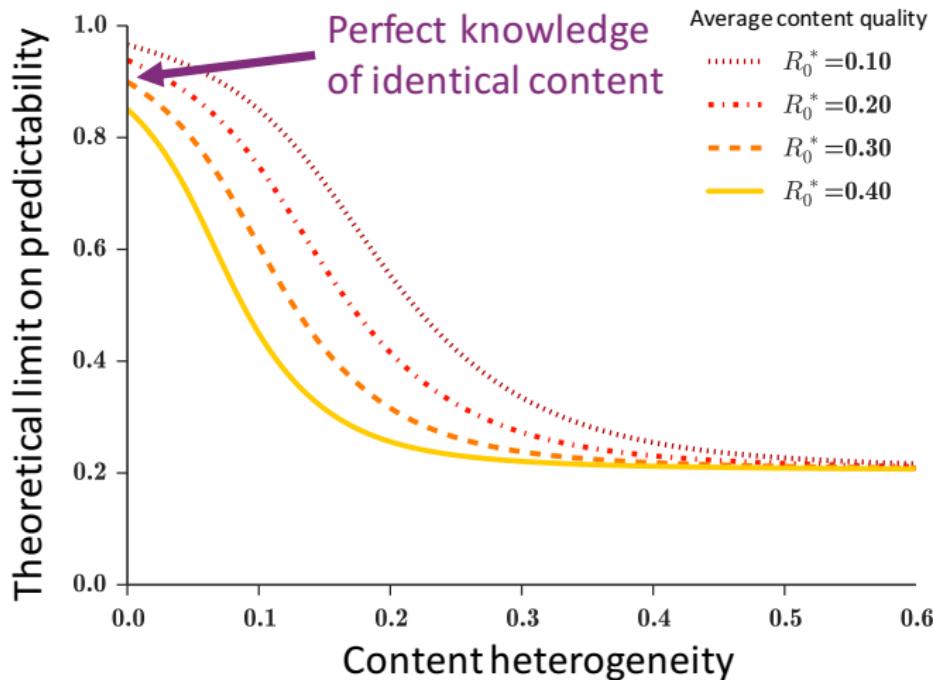
(c)



(d)

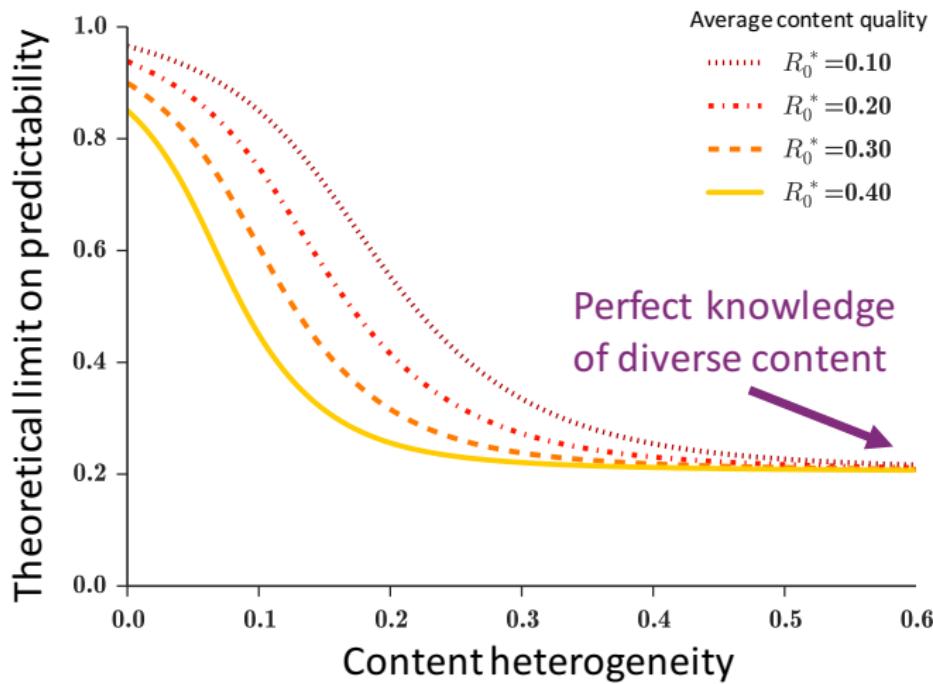
Repeatedly seed the same user with the same content

Outcomes are highly predictable when all content is identical



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%)



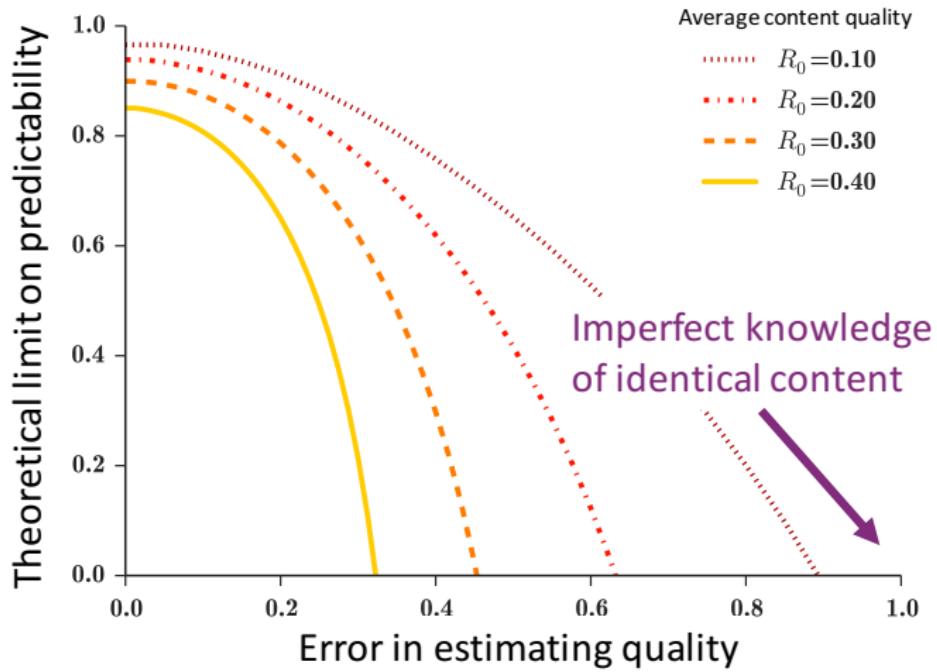
Repeatedly seed the same user with the same content

Outcomes are highly predictable assuming exact quality estimates



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

- 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
- 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

State-of-the-art models explain less than half of the variance in outcomes, based primarily on past success

Conclusions

This performance is likely due to randomness in diffusion process itself, rather than our ability to estimate or model it

Going forward

This is just one example of how digital platforms and computational thinking can advance our understanding of how people behave and interact

Going forward

Important to view prediction and explanation as compliments,
not substitutes

Computer science

 \hat{y}

Predict

vs
and

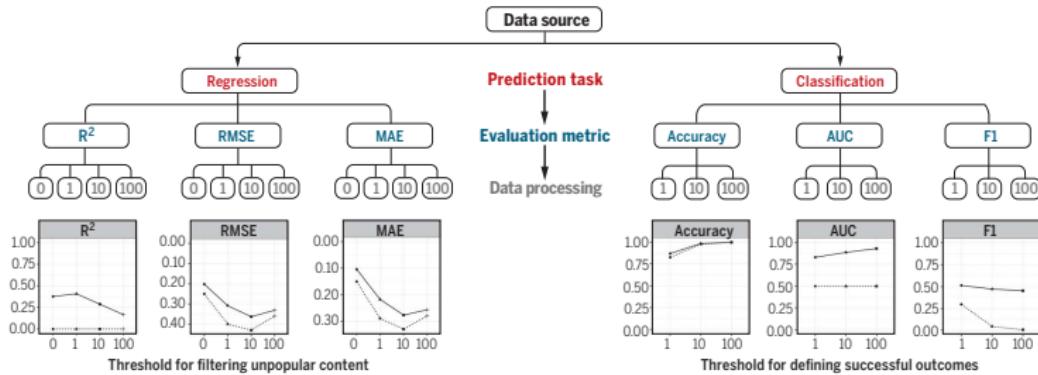
Social science

 $\hat{\beta}$

Explain

Going forward

Otherwise it can be difficult to make long-term progress in advancing social science



Going forward

Guidelines:

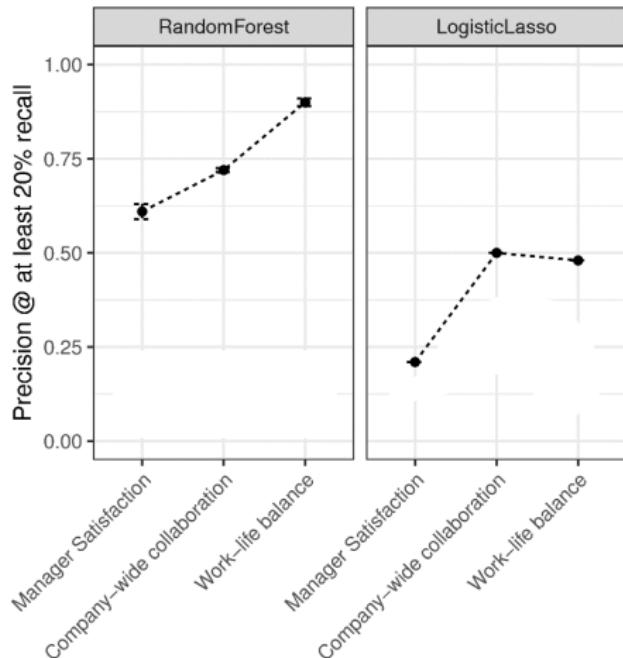
- Improved standards for confirmatory vs. exploratory research (common task framework, pre-registration)
- Focus on models that are both predictive and interpretable
- Entertain complex models for a complex world
- Consider inherent limits to predictability of social phenomena

Example: Predicting employee satisfaction

Predicting employee satisfaction on a yearly poll using internal communication data:

- I am satisfied with the level of company-wide collaboration
- I am satisfied with the effectiveness of my manager
- I am satisfied with my work-life balance

Results on the test set in year 1



Pre-registering predictive models

The screenshot shows two side-by-side project pages on the Open Science Framework.

Left Project Page:

- Title:** Prediction of employee attitudes using ...
- Contributors:** Amit Sharma
- Date registered:** 2016-10-10 10:27 AM
- Date created:** 2016-10-06 04:57 PM
- Category:** Project

Wiki Content:

Microsoft conducts an employee poll every year where it asks employees to answer questions about their work, their manager, and the company overall. Our goal is to predict employee responses to poll questions using metadata from employees (e.g., timestamps and anonymized recipient lists). We obtain anonymized records of email activity by employees. Using these and other non-email features.

Right Project Page:

- Title:** Prediction of employee attitudes using ...
- Description:** This registration is a frozen, non-editable version of [this project](#).
- File List:** hashed_predictions.txt (Version: 1), prereg_poll_predictors.R
- OSF Storage:** OSF Storage, Archive of OSF Storage

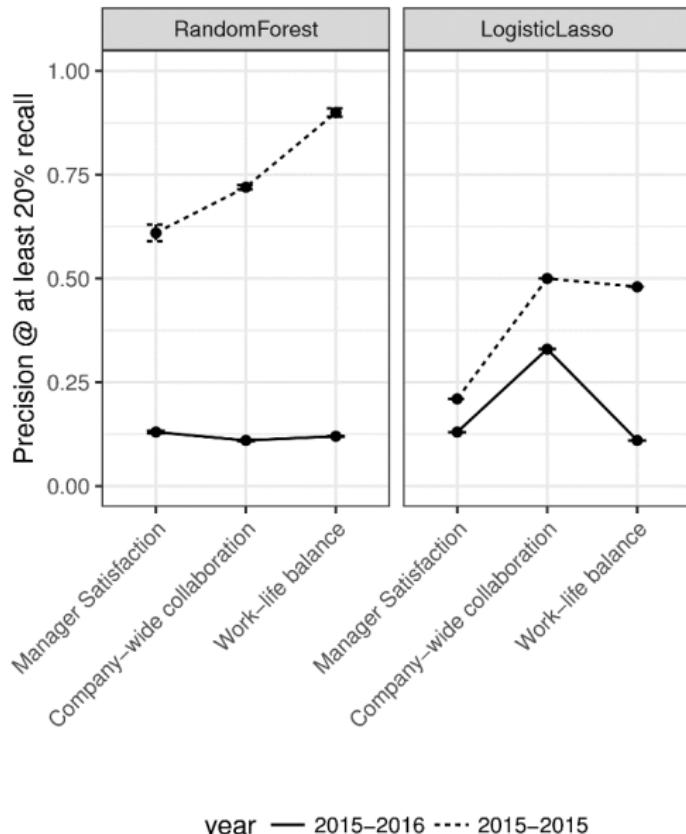
File Details for hashed_predictions.txt:

| Hash |
|-------------------------------------|
| bc3f268c2a64daa41d02d528e669012e |
| 31d640b49867fafb8a296ec8c2bd3ec |
| d389aed7750d4a1da6cc340f388209c |
| fc81478a627b57c9e0e2e1a220233e09 |
| 89ad825e906ce47db52ced9415f55aac |
| a14cd8a131c38d4d8e8ca3356dc05967f |
| d574ccdf7aff4f40b0cd1d103b84663a133 |
| ba68686b63288a7bc26629cdff5e928b |
| be9aa893582cef82d058a8cd2473778 |
| c42c2f751ace27f2e2d265b7b5bb7f |
| e919bd4c9296c5c9b967ae1f6d83ae51 |
| 2fccccb990d593c6dd119587da70fc7 |
| 9a7c7f6cb8d996daecfe260b41afcc |
| 60c52519da56d395a44fd8ed4569ec |
| 9997dbea42ec3b322764b42c9038911e |
| 8f5b013d87e24ac2645cb2086e68b41b |

Pre-registering predictive models

- Random seed used to initialize pseudo-random generation
- Procedure for dividing data into training, evaluation, test sets
- Outcome variable for prediction
- Features used by the model, along with any transformations
- Model class used with pre-specified hyperparameters
- Metric(s) used for evaluating the trained models predictions

Results on year 2: twice the work, none of the reward



Thanks. Questions?

jmh@microsoft.com
@jakehofman

“Exploring limits to prediction in complex social systems”

Travis Martin, Jake Hofman, Amit Sharma, Ashton Anderson, and Duncan Watts

WWW 2016

bit.ly/predlim

“Prediction and explanation in social systems”

Jake Hofman, Amit Sharma, and Duncan Watts

Science 2017

bit.ly/predictexplain