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

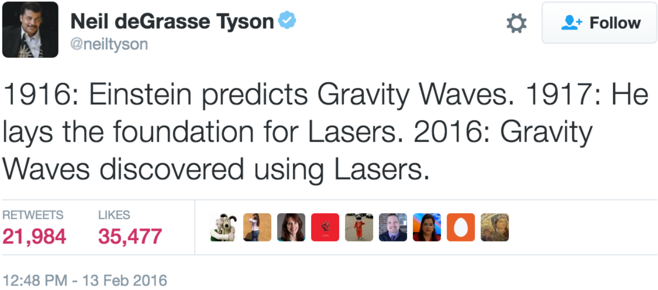
Follow

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

RETWEETS	LIKES
???	???

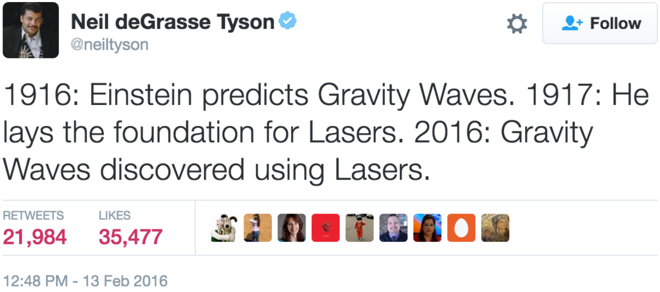
12:48 PM - 13 Feb 2016

How far will this spread?



Why is so difficult to predict success?

Do we need bigger data and better models?



Or is information diffusion inherently unpredictable?

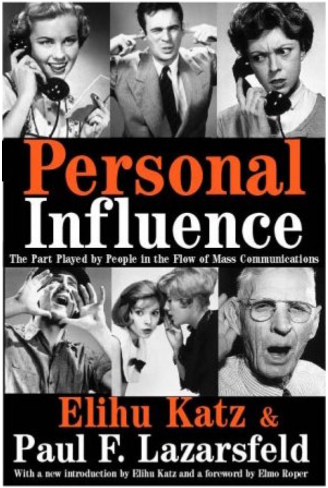
Outline

- 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

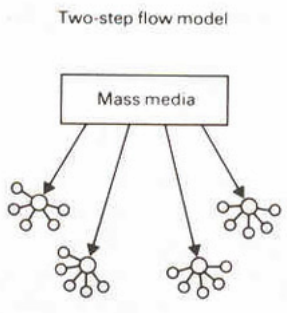


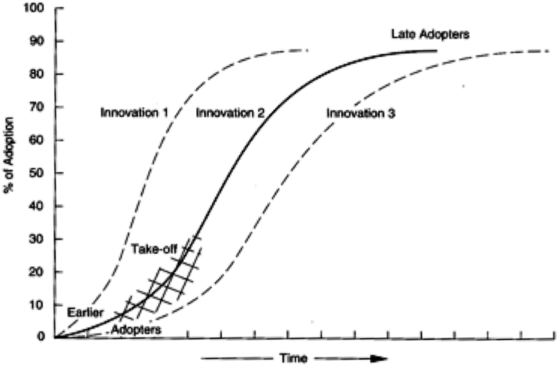
Table 56—The "Cosmopolitans" Among the Opinion Leaders Are in Fashions and Public Affairs

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)

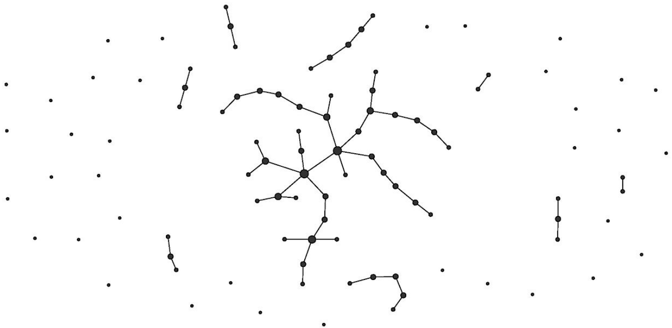
Katz & Lazarsfeld (1955)

∼1960s: Mathematical models of aggregate adoption



Rogers (1962), Bass (1969)

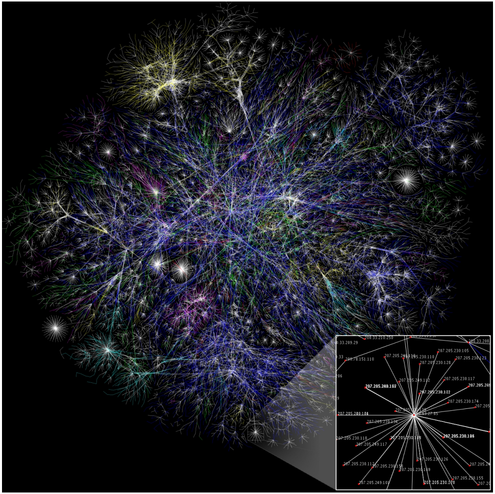
∼1960s: Random graph theory



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

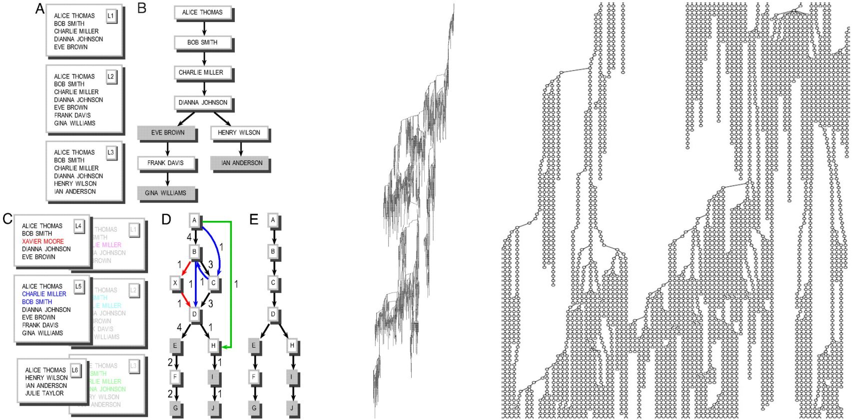
Erdős & Rényi (1959)

∼1990s: Empirical structure and dynamics of networks



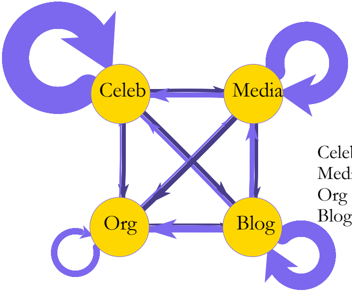
Newman, Barabasi, Watts (2006)

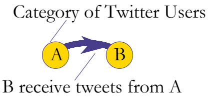
∼2000s: Empirical analyses of large-scale diffusion events



Liben-Nowell & Kleinberg (2007)

∼2010s: Characterizing online information flows

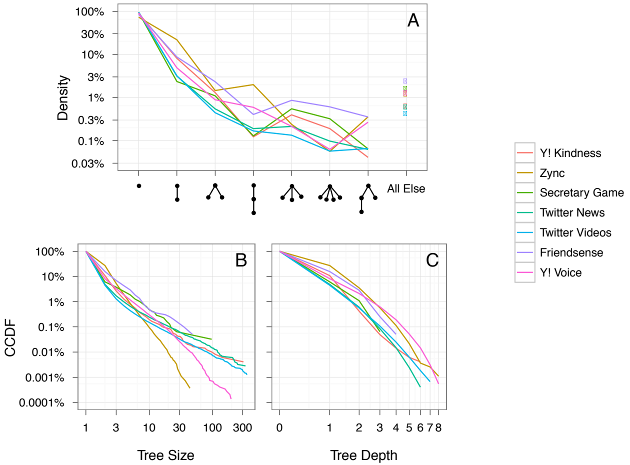




	% of tweets received from			
	Celeb	Media	Org	Blog
Celeb	38.27	6.23	1.55	3.98
Media	3.91	26.22	1.66	5.69
Org	4.64	6.41	8.05	8.70
Blog	4.94	3.89	1.58	22.55

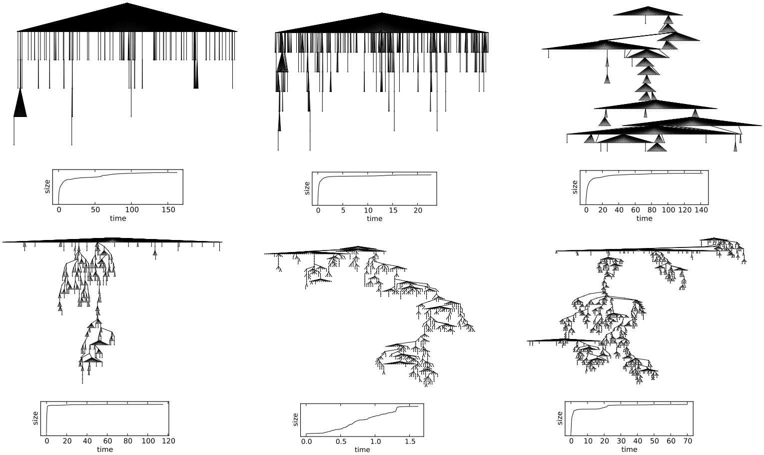
Wu, Hofman, Mason, Watts (2011)

∼2010s: Cataloging empirical diffusion structures



Goel, Goldstein, Watts (2012)

∼2010s: Cataloging empirical diffusion structures



Goel, Anderson, Hofman, Watts (2015)

2016

- 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

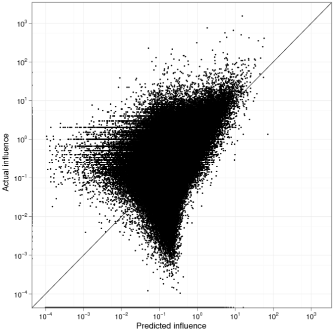
Predicting Success

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

Background: Predicting the success of diffusion events

Bakshy, Hofman, Mason, Watts (2011)

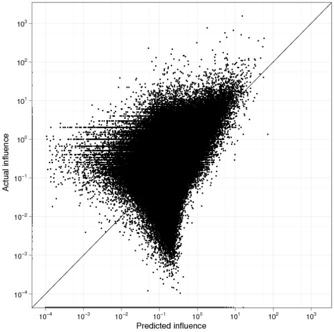
- Looked at 75M diffusion events across 1M users
- Found a relatively low correlation ($R^2 \sim 30\%$) between predicted and actual cascade sizes
- Almost all predictive power comes from examining past performance of a user or piece of content



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 (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

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[2]

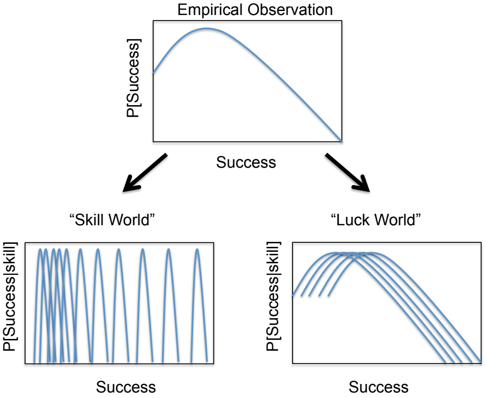
- Model success S as a mix of skill Q and luck ϵ:

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

- Measure the fraction of variance remaining after conditioning on skill:

$$F = \frac{\mathbb{E}[\mathrm{Var}(\mathrm{S}|Q)]}{\mathrm{Var}(\mathrm{S})} = 1 - R^2$$

- $R^2 = 1$ 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

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

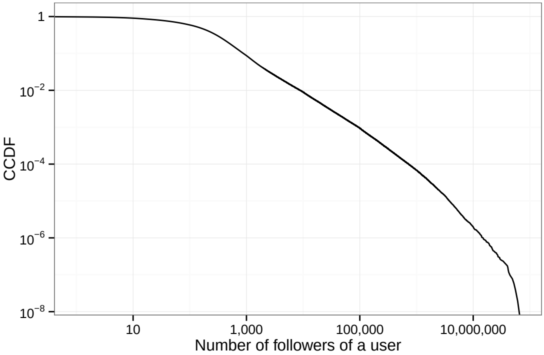
- Examined all 1.4B tweets containing URLs posted in February 2015
- Eliminated spam using internal Microsoft classifier
- Restricted attention to tweets containing URLs from the top 100 English-speaking domains with the most unique adopters
- Resulted in 850M tweets from 50M distinct users covering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

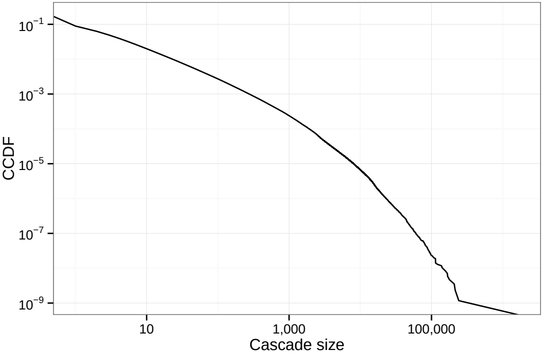
User distribution

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



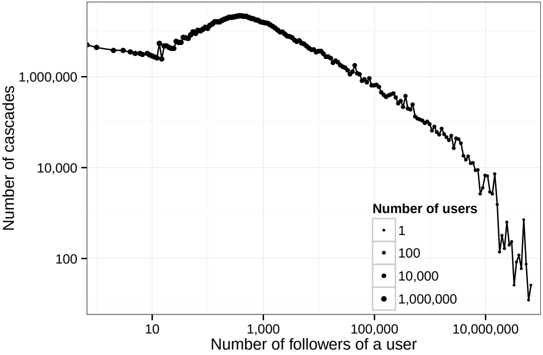
Cascade sizes

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



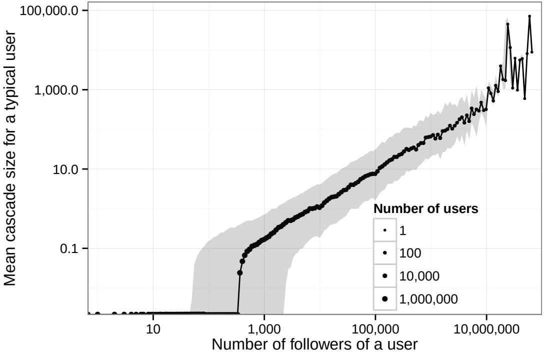
Activity by degree

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

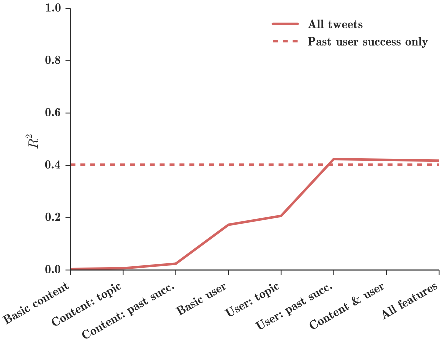
Used a random forest to estimate success (cascade size) given skill (available features)

- 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 half of the variance in outcomes

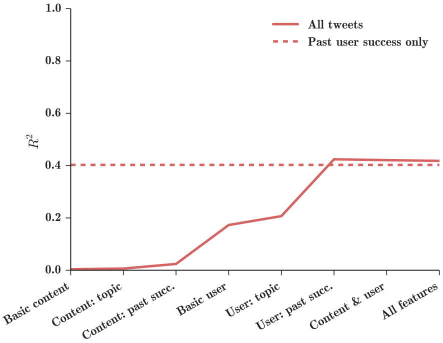
Model	Tweet time	Domain	Spam score	Category	Tweet topic	Past url success	User time	Followers	Friends	Statuses	User topic	Past user success	Topic interaction
1. Basic content	✓	✓	✓	✓									
2. Content, topic	✓	✓	✓	✓	✓								
3. Content, past succ.	✓	✓	✓	✓	✓	✓							
4. Basic user							✓	✓	✓	✓			
5.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



Predictive performance

Content features alone perform poorly

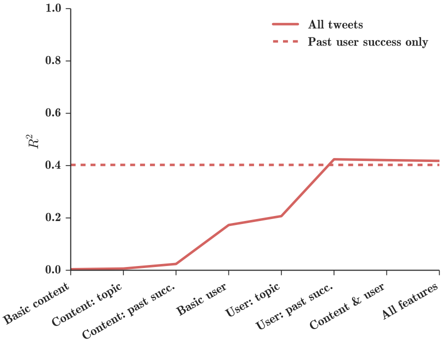
Model	Tweet time	Domain	Spam score	Category	Tweet topic	Past url success	User time	Followers	Friends	Statuses	User topic	Past user success	Topic interaction
1. Basic content	✓	✓	✓	✓									
2. Content, topic	✓	✓	✓	✓	✓								
3. Content, past succ.	✓	✓	✓	✓	✓	✓							
4. Basic user							✓	✓	✓	✓			
5.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



Predictive performance

Basic user features provide a reasonable boost in performance

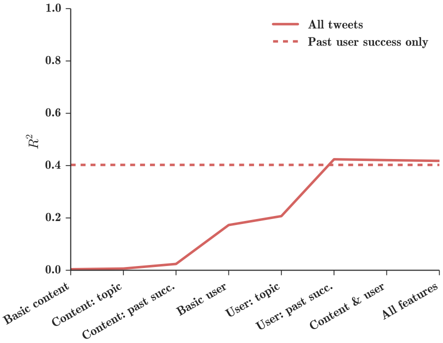
Model	*Tweet time*	*Domain*	*Spam score*	*Category*	*Tweet topic*	*Past url success*	*User time*	*Followers*	*Friends*	*Statuses*	*User topic*	*Past user success*	*Topic interaction*
1. Basic content	✓	✓	✓	✓									
2. Content, topic	✓	✓	✓	✓	✓								
3. Content, past succ.	✓	✓	✓	✓	✓	✓							
4. Basic 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	Tweet time	Domain	Spam score	Category	Tweet topic	Past url success	User time	Followers	Friends	Statuses	User topic	Past user success	Topic interaction
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 $R^2 \sim 50\%$
- 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

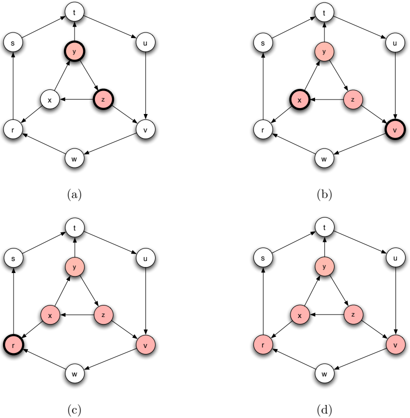
(Exploring the limits to predicting success)

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

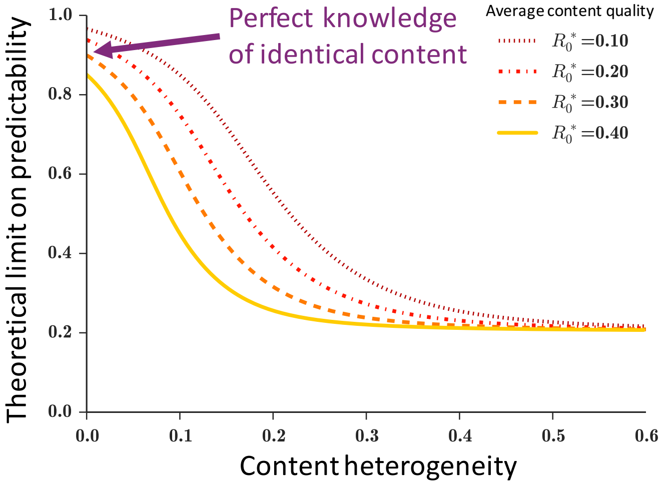
Simulations

- 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



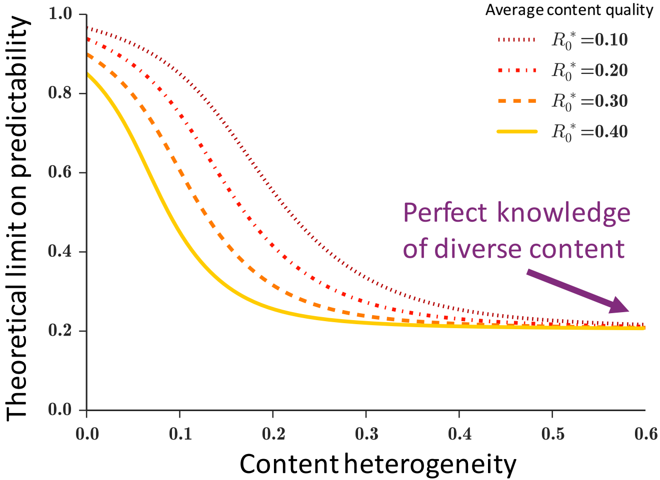
Repeatedly seed the same user with the same content

Outcomes are highly predictable when all content is identical



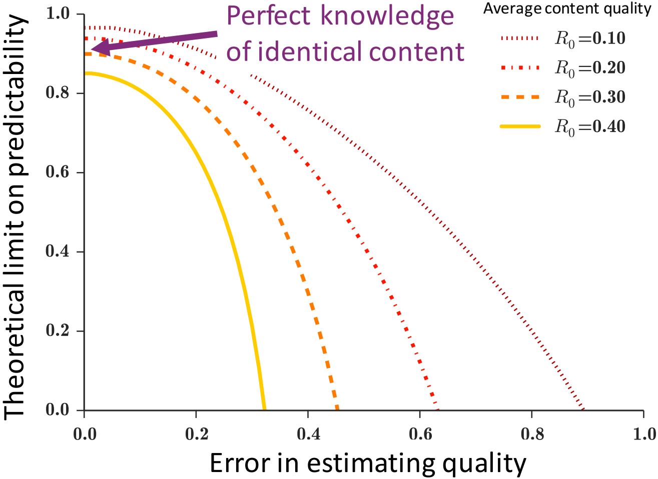
Repeatedly seed the same user with the same content

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



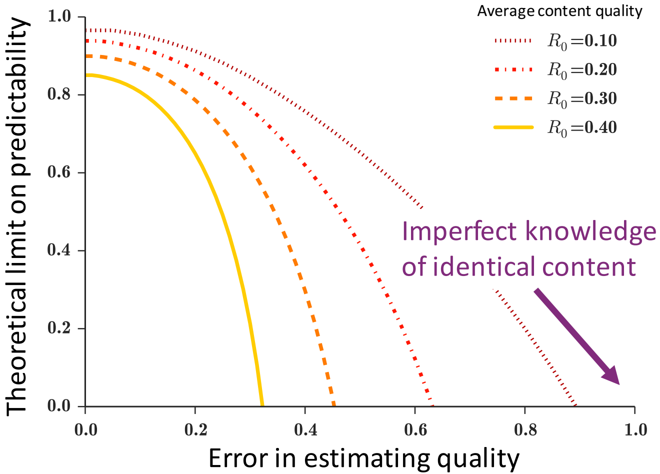
Repeatedly seed the same user with the same content

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

- 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

Conclusions

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

Conclusions

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