Exploring limits to prediction in complex social systems: Predicting cascade size on Twitter

Travis Martin

Jake Hofman, Amit Sharma

Ashton Anderson, Duncan Watts

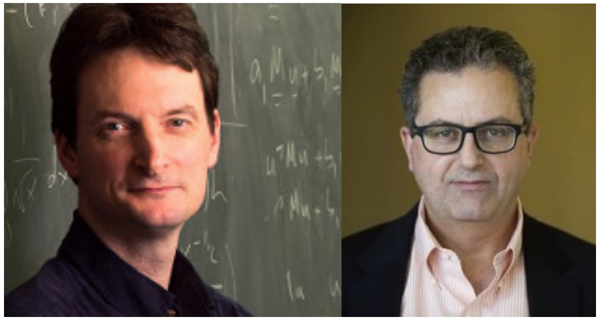
Microsoft
Research

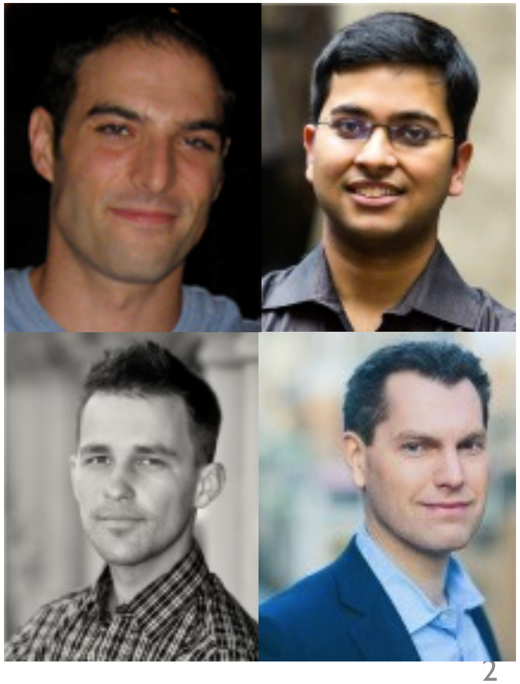
A personal introduction

University of Michigan,

Computer Science

- Network science





Summer @ Microsoft Research

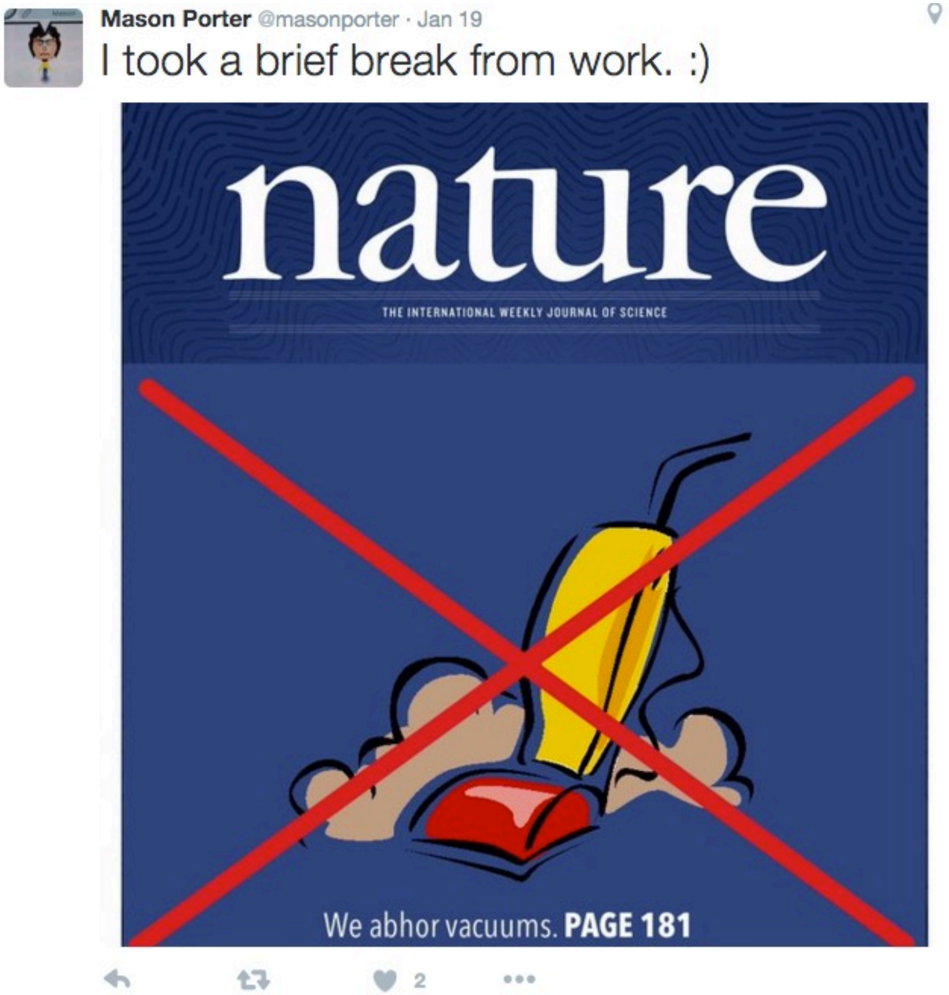
- Early work on *hard* problem
- Please ask me questions
- WWW 2016

Predicting success on Twitter?

Bakshy, Hofman, Mason, Watts (2011):

How viral will my tweet be?

“Cascades are unpredictable!”



Mason Porter @masonporter · Jan 19

I took a brief break from work. :)

nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
We abhor vacuums. PAGE 181
2

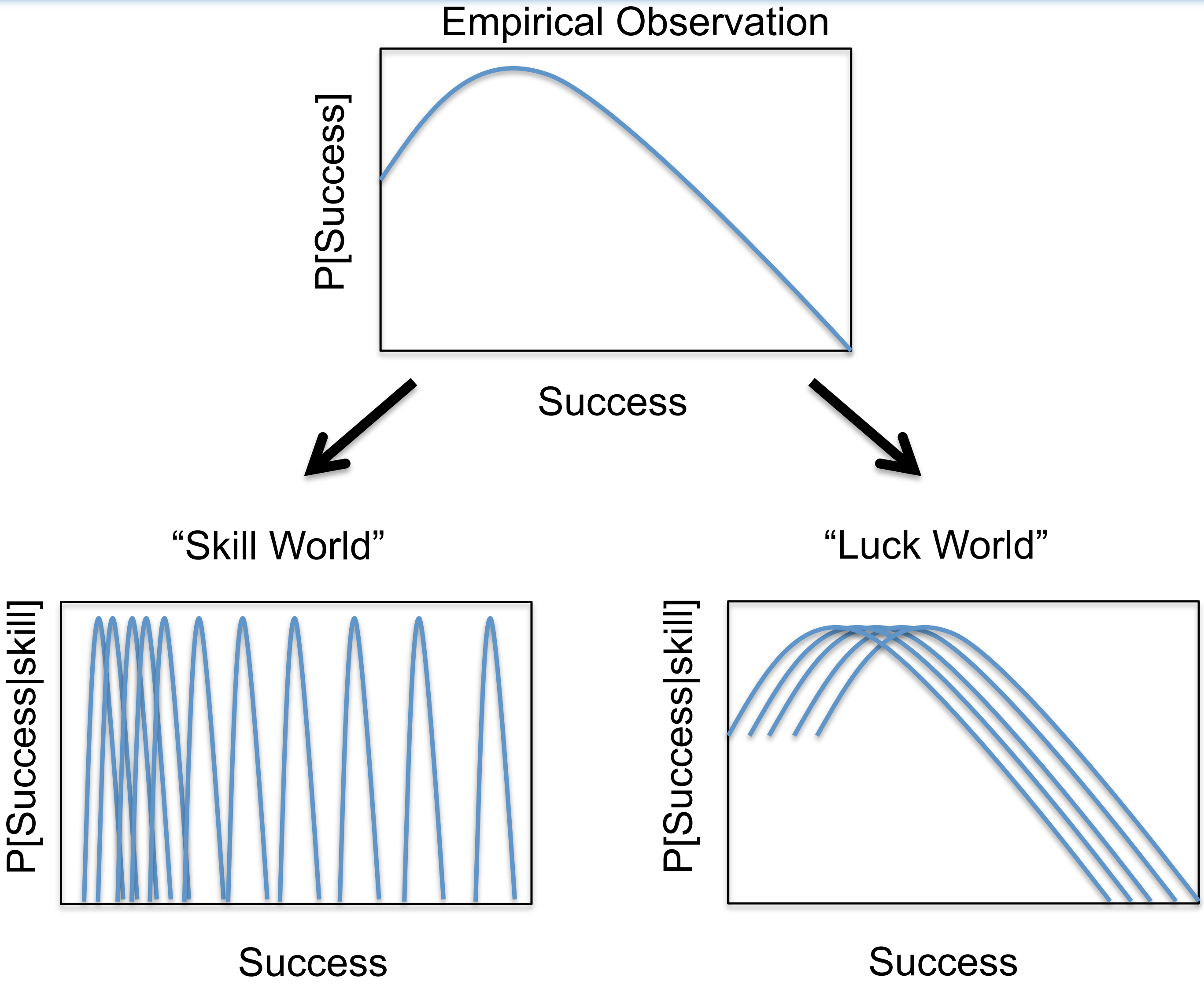
Incomplete history of cascade prediction

Who	Predicting	Features	Metric	Conclusion
HongD 10	Is item retweeted?	Topic Models	F1=0.47	Better than baseline
JendersKN 13	Will item reach some size T?	Content	F1>0.9	High accuracy
TanLP 14	Which of two does better?	Wording	Accu=65.6%	Computers are OK
ChengADKL 14	Will cascade double?	Temporal	AUC=0.88	Predictable
Lerman, Yang, Petrovic, Romero, Kupavskii, Ma, Weng, Zhao, Yu, etc				

‘Predictable’ needs a definition

1. A framework for predictability
2. Explore the predictability of information cascades (Twitter) within this framework
3. Simulation results
4. Future ideas for measuring predictability

Distinguishing model error from randomness



Empirical Observation

“Skill World”

“Luck World”

Unpredictable: imperfect prediction with perfect model

Our two approaches for information cascades:

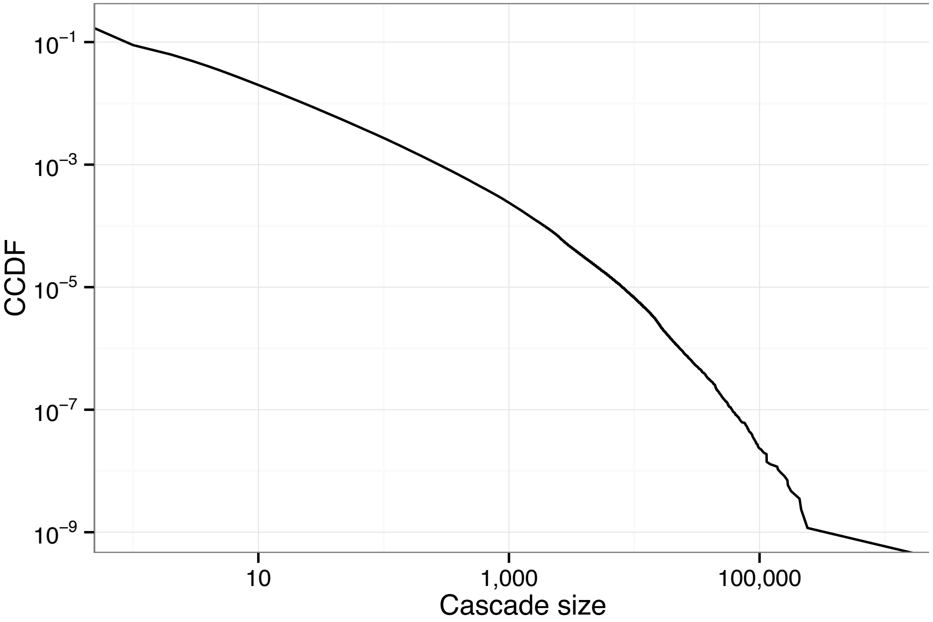
1. (Empirical) Does prediction performance plateau with better models and data?
2. (Simulation) Is performance highly sensitive to noise?

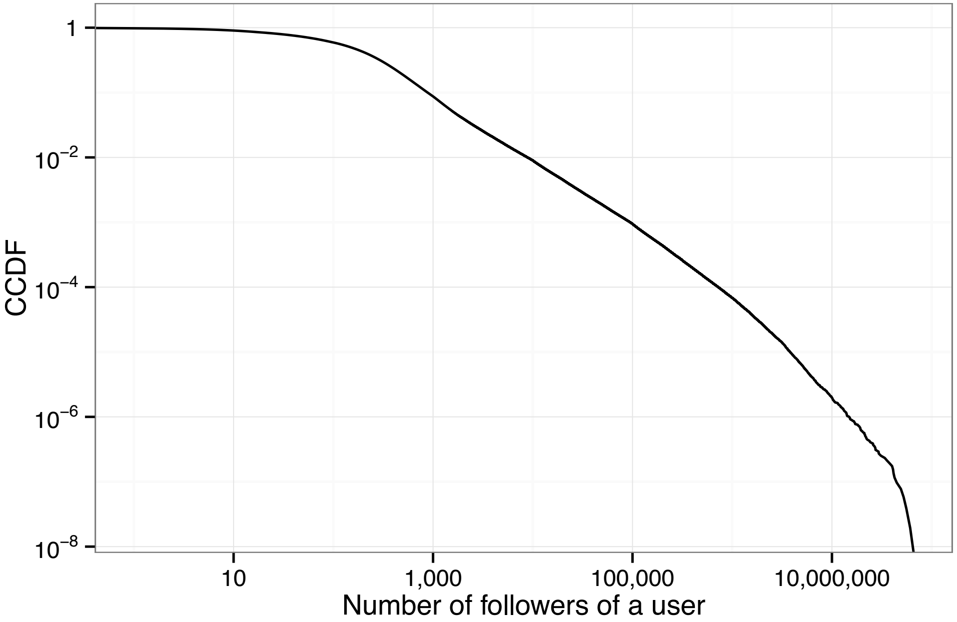
Why Twitter

- If we can't predict things on Twitter, can we in the real world?
  - Lots of data
  - Fully observable spread
- Information cascades

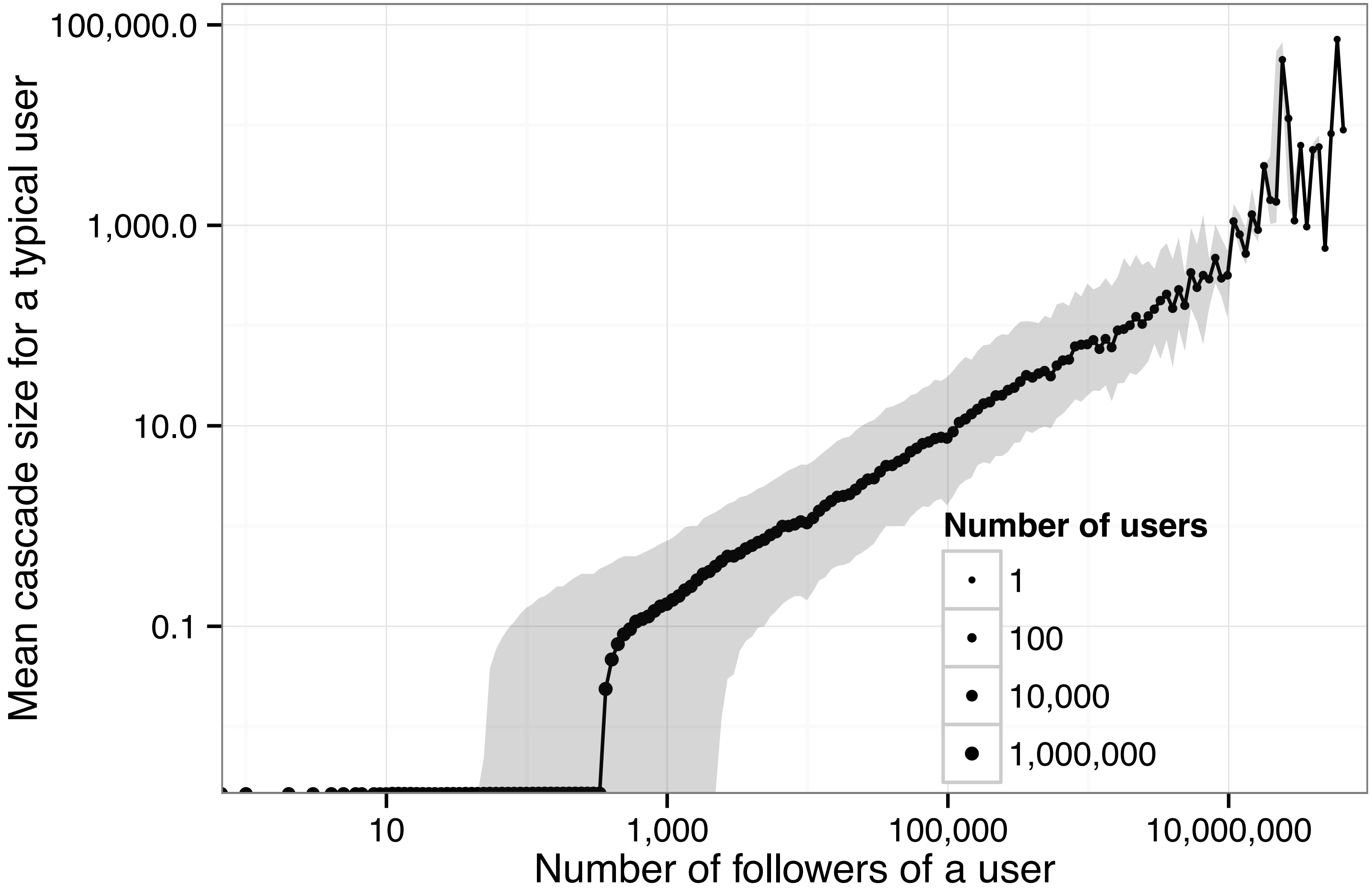
Cascade size

Followers

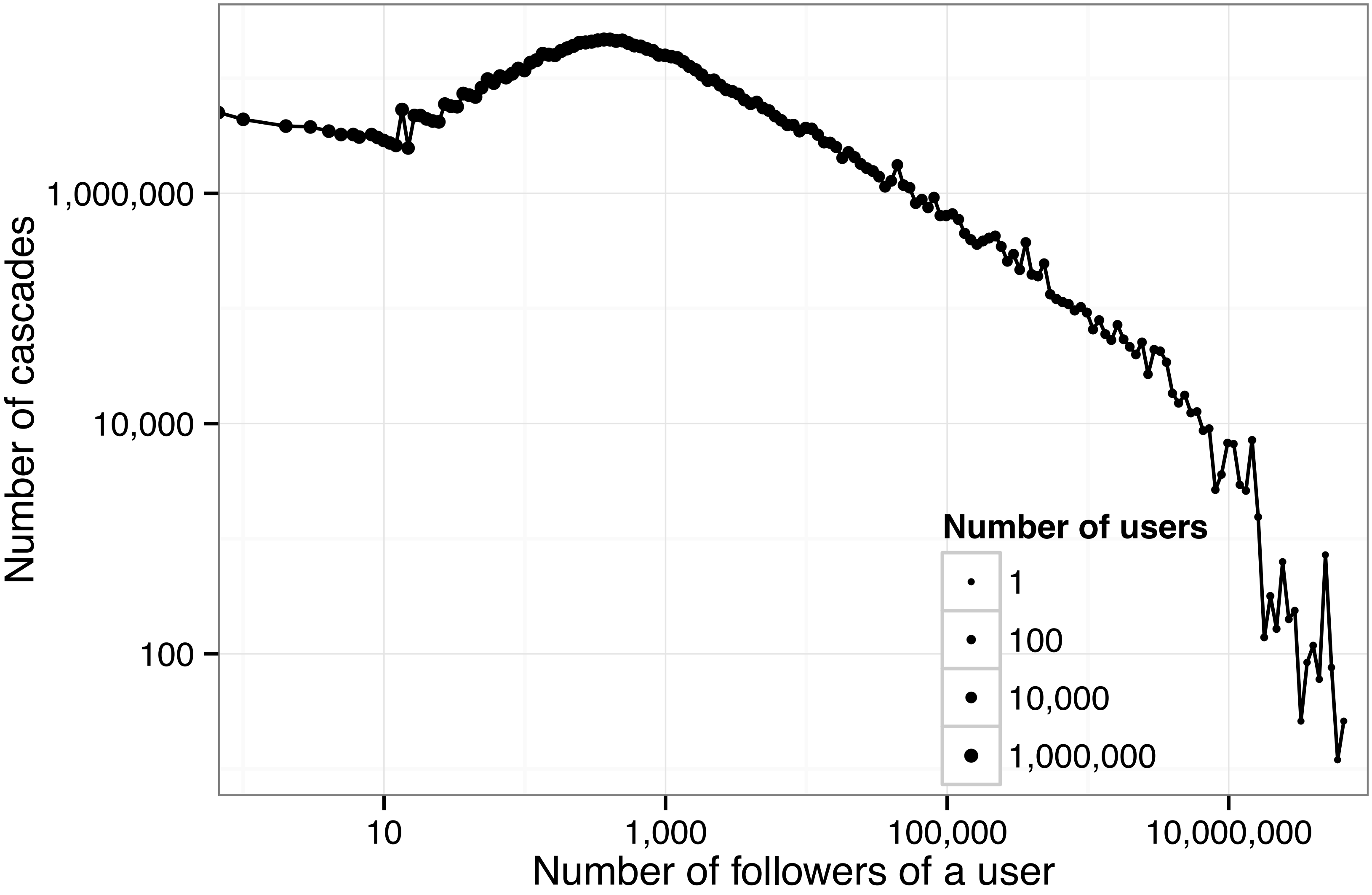




Cascade size vs degree

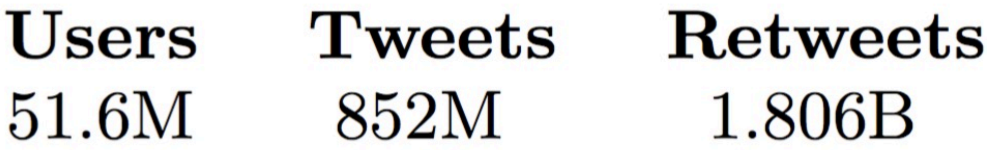


# tweets vs degree



Our task

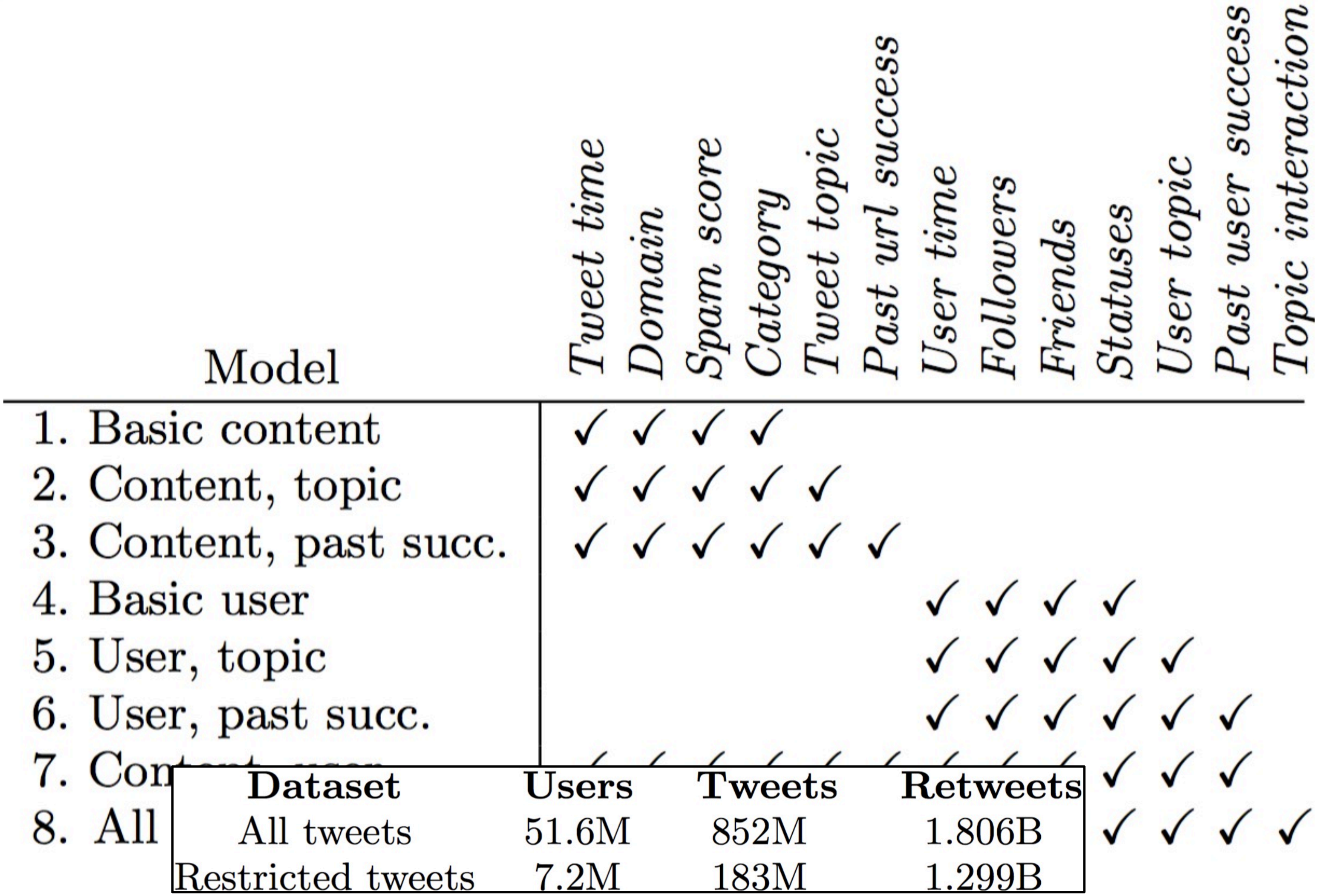
- Predict final # retweets of tweets with urls
- Filter to 100 popular domains
- February 2015:



Users	Tweets	Retweets
51.6M	852M	1.806B

- Features:
  - Tweet information
  - User information
- Optimize R^2
  - (MSE, reduction in variance)

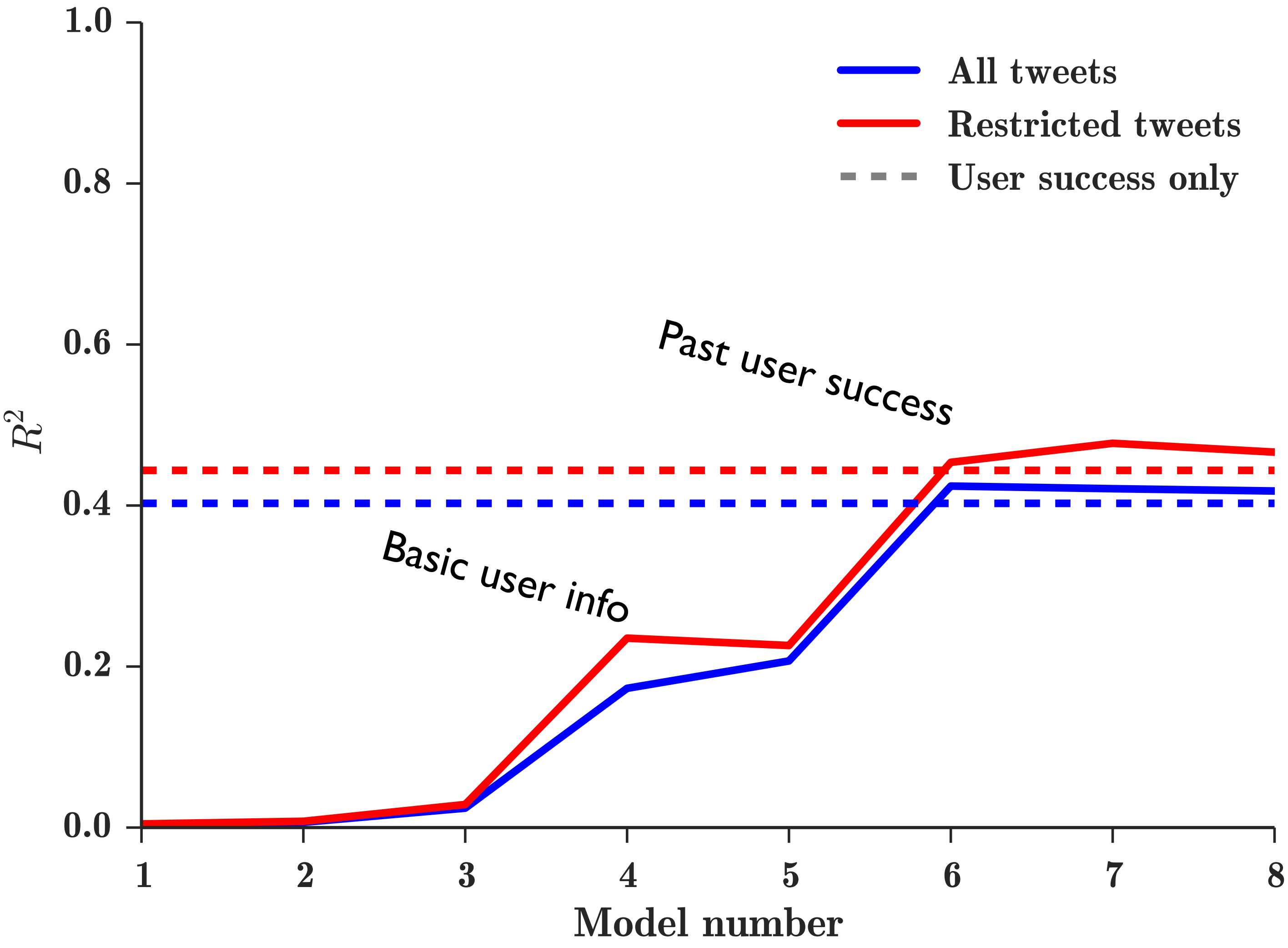
Random forest features



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										✓	✓	✓	✓

Dataset	Users	Tweets	Retweets
All tweets	51.6M	852M	1.806B
Restricted tweets	7.2M	183M	1.299B

Prediction limit on twitter



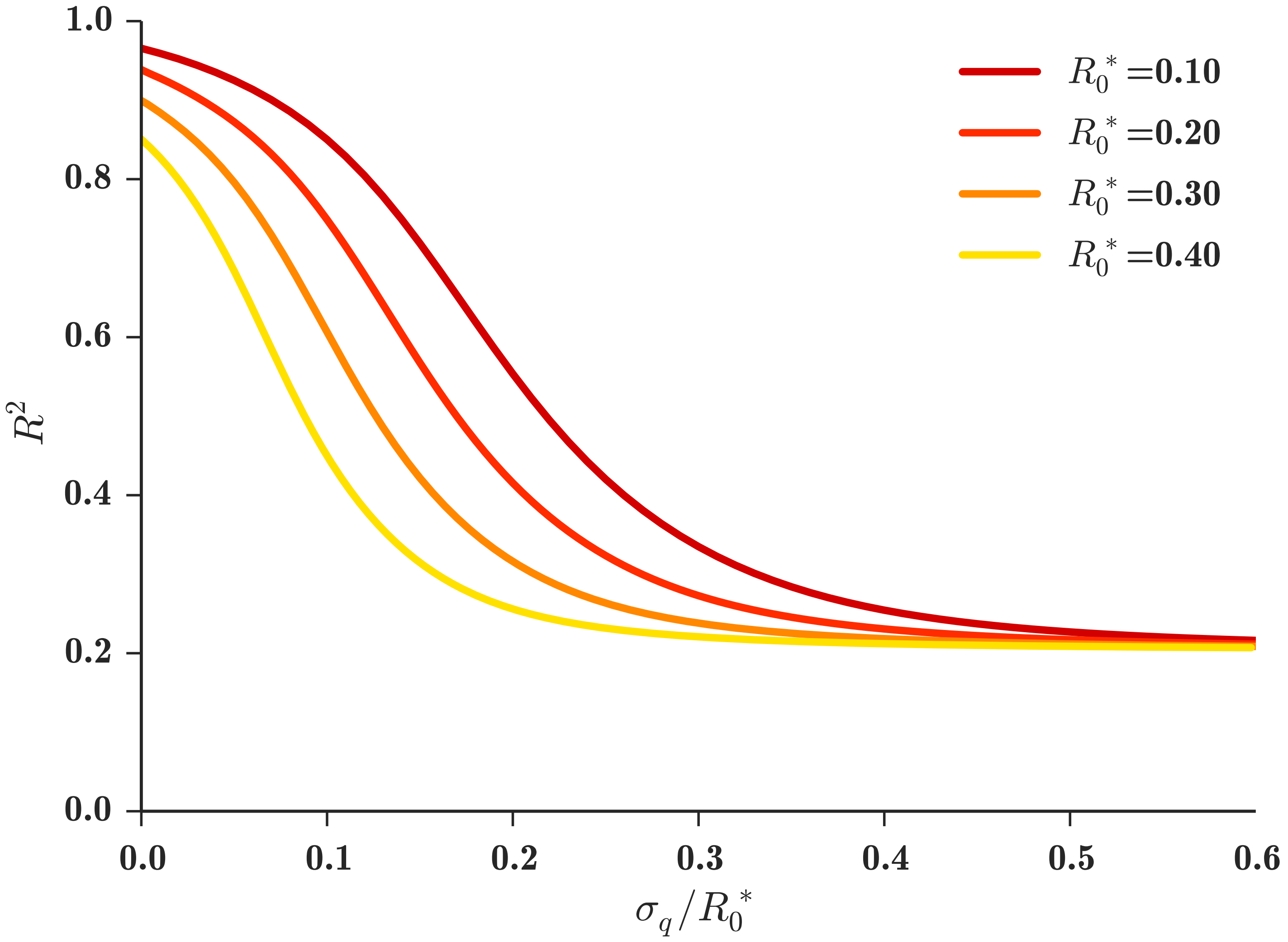
How can you *prove* a limit?

- Results robust to other ML models
  - Decision tree, linear regression
- Consistent with prior work
- Asymptote, dependency between features
- Can’t rule everything out
  - Simulation

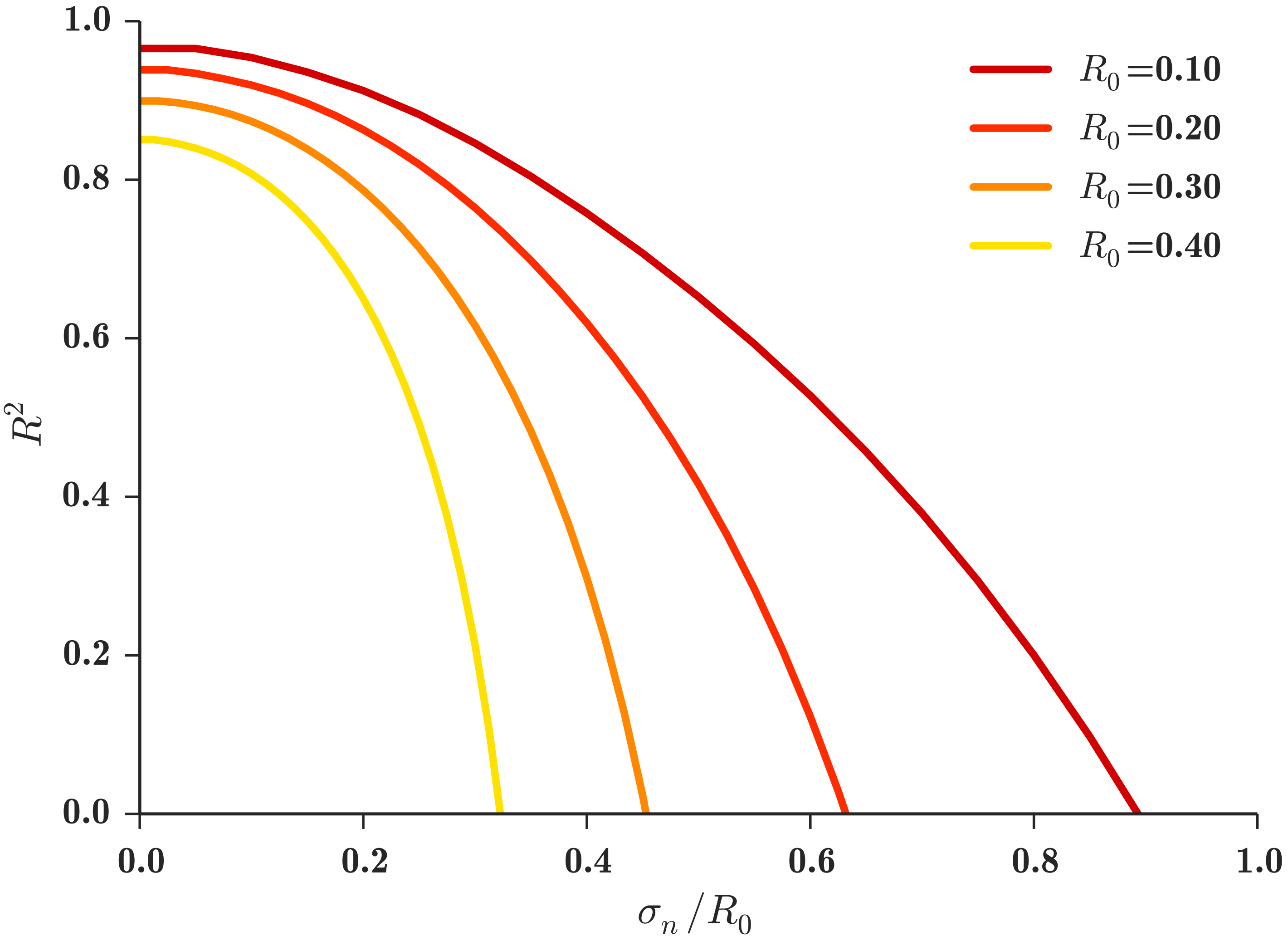
Simulation

- SIR disease model
- Scale free network similar to Twitter
  - 7M users, $\alpha = 2.05$
  - 8B simulated cascades
- *Quality*: R_0 = average neighbors infected
  - *p*(infect over edge) x *mean-degree*
- Prediction task
  - Given (possibly noisy) estimate of R_0 and the seed node, predict cascade size

Increasingly heterogeneous quality



Increasing noise



Conclusion

1. Unifying framework for skill vs luck
2. Most extensive study of Twitter
   - Apparent limit to prediction
3. Simulation shows sensitivity to noise, heterogeneity

More ideas

1. In some cases randomness averages out
   – How/why are cascades different?
2. Are there any controlled or natural experiments we can do?
3. Better measurements of prediction goodness
   – R^2 is sensitive to outliers
4. More features, time dependence
   – How independent are Twitter features?
5. More realistic simulation models

Thanks!

travisbm@umich.edu
arxiv.org/abs/1602.01013
travismart.com